
TEACHING INTRODUCTORY PHYSICS

ARNOLD B. ARONS

*Teaching Introductory
Physics*

Arnold B. Arons

UNIVERSITY OF WASHINGTON



JOHN WILEY & SONS, INC.

New York Chichester · Brisbane · Toronto Singapore

ACQUISITIONS EDITOR Stuart Johnson
MARKETING MANAGER Catherine Faduska
PRODUCTION EDITOR Deborah Herbert
MANUFACTURING MANAGER Mark Cirillo

This book was set in 10 pt Times Roman by the author and
printed and bound by Hamilton Printing Company. The cover was printed by
Hamilton Printing Company.

Recognizing the importance of preserving what has been written, it is a
policy of John Wiley & Sons, Inc. to have books of enduring value published
in the United States printed on acid-free paper, and we exert our best efforts to that end.

Copyright © 1997, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of
this work beyond that permitted by Sections
107 and 108 of the 1976 United States Copyright
Act without the permission of the copyright
owner is unlawful. Requests for permission
or further information should be addressed to
the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Arons, A. B. (Arnold B.)

Teaching introductory physics / Arnold B. Arons.

p. cm.

Includes bibliographical references and index.

ISBN 0-471-13707-3 (alk. paper)

1. Physics--Study and teaching. I. Title.

QC30.A78 1996

530'.071'1--dc20

96-16838

CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PART I

A Guide to Teaching for Learning and Understanding

Preface to Part I

Starting approximately twenty years ago, members of the physics teaching community began conducting systematic observations and research on student learning and understanding of physical concepts, models, and lines of reasoning. Some of these investigations began with, or subsequently spilled over into, research on more general aspects of the development of the capacity for abstract logical reasoning. In Part I of this book, I have tried to bring together as many as possible of the relevant insights into the teaching of the most basic aspects of introductory physics—covering high school through first year college level, including basic aspects of the course aimed at physics and engineering majors, without penetrating the full depth of the latter.

Very little that I present is based on conjecture. I have invoked and referred to most of the systematic research of which I am aware, and I have drawn on my own observations, which have been under way for more than forty years and have been extensively replicated over that time. One of my sources has been the direct interview in which one asks questions and listens to the individual student response; the other has been the analysis of students' written response to questions on tests and examinations. It is impossible to give all of the protocols of student interviews and all of the detailed supporting evidence without producing a book of impossible length. Although I give specific examples of student response from time to time, some of the insights are asserted without the full support they deserve. I can only ask the careful and critical reader to bear with these gaps, test them as opportunity arises, or turn to the more detailed literature for deeper penetration .

It is also impossible to include, in a book of reasonable length, all of the insights emerging from research on teaching, learning, and cognitive development. The literature is rich, varied, and rapidly increasing. I have been selective and have tried to include observations having the most direct bearing on classroom practice at the most basic levels of subject matter; the list of references will open the door to those wishing to pursue greater detail and explore primary evidence. Where a significant reference at this level is missing, the fault is in my judgment or in my not having fully encompassed the

extensive literature.

Both the *American Journal of Physics* and *The Physics Teacher* are rich in articles discussing the logic and epistemology of various laws and concepts, outlining improved modes of presentation, suggesting demonstrations and other ways of making abstractions clearer and more concrete, describing ways of engaging students in direct activities, criticizing loose and faulty approaches, introducing new derivations, new laboratory experiments, and so forth. Every one of these functions is valuable and important to our community, and I wish someone, more competent than I, would undertake to bring together the heritage that has accumulated over the years in these areas into another book on physics teaching.

It is necessary for me to make clear, however, that my own purpose is different. I have undertaken to discuss some of the elements that I believe underlie and precede a great many of the ideas and presentations appearing in the journals. In fact, many of the excellent suggestions appearing in the journals turn out to be ineffective with large numbers of students, not because of anything wrong with the suggestions, but because the students have not had a chance to master the necessary prior concepts and lines of abstract logical reasoning. It is to this end that I have elected to concentrate on some of these prior aspects of cognitive development and on underlying problems of learning and understanding that have been commanding increasing attention in recent years. In doing this, I in no way disparage the valuable materials and modes of presentation that are described in the journals and that enter in full force at the points where I leave off.

It must further be emphasized that I am not formulating prescriptions as to how items of subject matter should be presented to the students or how they should be taught, nor am I suggesting that there is one single way of getting any particular item "across to the student." There is tremendous diversity in style and method of approach among teachers, and such diversity should flourish. My objective is to bring out as clearly and explicitly as possible the conceptual and reasoning difficulties many students encounter and to point up aspects of logical structure and development that may not be handled clearly or well in substantial segments of textbook literature. With respect to modes of attack on these instructional problems (avenues of explanation and presentation, balance of laboratory versus classroom experience, use of computers and of audiovisual aids), I defer to the style and predilections of the individual teacher.

I have endeavored to cover the range from high school physics through college and university calculus-based courses. Some of the material, therefore, goes well beyond high school level, and high school teachers should draw appropriate lines, limiting the more sophisticated material to their front running students if invoking it at all. At the other end of the spectrum, teachers in colleges enrolling highly selected students, or teachers with a highly selected student body in calculus-based engineering-physics courses will find less rele-

vance in the discussions of some of the more mundane underpinnings. However, it is necessary to issue a warning: there is much more overlap between the disparate populations than most teachers realize, and it is frequently startling to find how many students, at a presumably fairly high level, have the same difficulties, preconceptions, and misconceptions as do much less sophisticated students. It is only the *fraction* of students having a certain difficulty that changes continuously as one goes up or down the scale; there is not an abrupt drop to zero at some intermediate level. Also, students at higher levels of scholastic ability, especially verbal skills, can usually remediate or overcome such initial difficulties at a more rapid pace than do other students, and a teacher needs to calibrate each of the classes with which he or she must deal.

Some of the chapters in Part I contain end-sections giving illustrations of possible test questions or homework problems. To keep down discursive length, I have not included detailed discussions of these questions and have only inserted occasional cryptic remarks about point and purpose. All these questions, however, are designed to implement some of the knowledge gained in the research protocols. They illustrate the kinds of questions that might be *added* to the normal regimen of quantitative end-of-chapter problems to confront the mind of the learner with aspects otherwise not being made explicit. Part II contains a more extensive collection of suggested homework and test questions of this variety. The examples being given in both Parts I and II are an invitation more than an end point. The pool of such questions must be greatly expanded to enhance variety and flexibility. Such expansion will take place not through the output of one individual, whose imagination gives out at some finite point, but through the superposition of effort on the part of numerous interested individuals, each of whom brings a new imagination to the effort. I long to see my limited set of examples greatly expanded.

Finally I point to the following unwelcome truth: much as we might dislike the implications, research is showing that didactic exposition of abstract ideas and lines of reasoning (however engaging and lucid we might try to make them) to passive listeners yields pathetically thin results in learning and understanding except in the very small percentage of students who are specially gifted in the field. Even in the calculus-based course, many students have the difficulties, and need all of the help, outlined in these pages. In expressing this caveat, I am, of course, *not* advocating *unclear* exposition. I am pointing to the necessity of supplementing lucid exposition with exercises that engage the mind of the learner and extract explanation and interpretation in his or her own words.

It is obvious that ideas and information such as I have summarized here cannot be developed in seclusion. I am deeply indebted to the hundreds of students who have submitted to my questioning, accepting the tension that goes with my shutting up and waiting for their answers. I am indebted also to the many colleagues and associates with whom I have discussed physics, prepared test questions, and worried about the meaning of learning and understanding.

standing. Among these are my former colleagues at Amherst College: the late Bruce Benson, Colby Dempsey, Joel Gordon, Robert Romer, the late Theodore Soller, and Dudley Towne; at the University of Washington: David Bodansky, Kenneth Clark, Ronald Geballe, James Gerhart, Patricia Heller, Lillian McDermott, James Minstrell, and the late Phillip Peters. Robert Romer and Kenneth Clark have read sections of Part I and have supplied me with valued criticism, corrections, and suggestions. Phillip Peters read all of it, and his substantive comments and advice were of immeasurable value and assistance.

Contents of Part I

CHAPTER 1	UNDERPINNINGS	1
1.1	Introduction	1
1.2	Area	1
1.3	Exercises with “Area”	2
1.4	Volume	3
1.5	Mastery of Concepts	3
1.6	Ratios and Division	4
1.7	Verbal Interpretation of Ratios	4
1.8	Exercises in Verbal Interpretation	5
1.9	Comment on Verbal Exercises	7
1.10	Arithmetical Reasoning Involving Division	8
1.11	Graphs and Arithmetical Reasoning	9
1.12	Scaling and Ratio Reasoning	12
1.13	Elementary Trigonometry	15
1.14	Horizontal, Vertical, North, South, Noon, Midnight	16
1.15	Interpretation of Simple Algebraic Statements	17
1.16	Language	18
1.17	Why Bother with Underpinnings	20
1.18	Examples of Homework and Test Questions	21
CHAPTER 2	RECTILINEAR KINEMATICS	23
2.1	Introduction	23
2.2	Misleading Equations and Terminology	23
2.3	Events: Positions and Clock Readings	25
2.4	Instantaneous Position	26
2.5	Introduction to the Concept of “Average Velocity”	26
2.6	Graphs of Position versus Clock Reading	28
2.7	Instantaneous Velocity	30
2.8	Algebraic Signs	32
2.9	Acceleration	32
2.10	Graphs of Velocity versus Clock Reading	35
2.11	Areas	36

2.12	Top of the Flight	37
2.13	Solving Kinematics Problems	38
2.14	Use of Computers	39
2.15	Research on Velocity Concept	40
2.16	Research on Acceleration Concept	42
2.17	Implications of the Research Results	45
2.18	Galileo and the Birth of Modern Science	46
2.19	Observation and Inference	50
2.20	Examples of Homework and Test Questions	51

CHAPTER 3 ELEMENTARY DYNAMICS 56

3.1	Introduction	56
3.2	Logical Structure of the Laws of Motion	57
3.3	An Operational Interpretation of the First Law	59
3.4	A Numerical Scale of Force	60
3.5	Inertial Mass	62
3.6	Superposition of Masses and Forces	64
3.7	Textbook Presentations of the Second Law	65
3.8	Weight and Mass	66
3.9	Gravitational versus Inertial Mass	67
3.10	Understanding the Law of Inertia	69
3.11	Some Linguistic Problems	73
3.12	The Third Law and Free-Body Diagrams	74
3.13	Logical Status of the Third Law	78
3.14	Distributed Forces	80
3.15	Different Arrows for Different Concepts	80
3.16	Understanding Gravitational Effects	81
3.17	Strings and Tension	88
3.18	“Massless” Strings	89
3.19	The “Normal” Force at an Interface	90
3.20	Accelerated Objects not “Thrown Backwards”	92
3.21	Friction	94
3.22	Demonstrations of “Inertia”	96
3.23	Different Kinds of “Equalities”	97
3.24	Solving Problems	99
3.25	Sample Homework and Test Questions	101

CHAPTER 4 MOTION IN TWO DIMENSIONS 107

4.1	Vectors and Vector Arithmetic	107
4.2	Defining a “Vector”	108
4.3	Components of Vectors	109
4.4	Projectile Motion	111

4.5	Phenomenological Thinking and Reasoning	114
4.6	Radian Measure and π	116
4.7	Rotational Kinematics	118
4.8	Preconceptions Regarding Circular Motion	119
4.9	Centripetal Force Exerted by Colinear Forces	121
4.10	Non-Colinear Forces	124
4.11	Frames of Reference and Fictitious Forces	127
4.12	The Two-Body Problem	128
4.13	Torque	131
4.14	Sample Homework and Test Questions	134
 CHAPTER 5 MOMENTUM AND ENERGY		135
5.1	Introduction	135
5.2	Developing the Vocabulary	136
5.3	Describing Everyday Phenomena	137
5.4	Force and Rate of Change of Linear Momentum	138
5.5	Heat and Temperature	139
5.6	Impulse-Momentum and Work-Kinetic Energy Theorems ..	142
5.7	Real Work and Pseudowork	145
5.8	The Law of Conservation of Energy	146
5.9	Digression Concerning Enthalpy	148
5.10	Work and Heat in Sliding Friction	150
5.11	Deformable System with Zero-Work Force	153
5.12	Rolling Down an Inclined Plane	154
5.13	Inelastic Collision	157
5.14	Some Illuminating Exercises	158
5.15	Spiralling Back	161
5.16	Sample Homework and Test Questions	163
 CHAPTER 6 STATIC ELECTRICITY		167
6.1	Introduction	167
6.2	Distinguishing Electric, Magnetic, and Gravitational Interactions	168
6.3	Electric Charge	169
6.4	Electrostatics Experiments at Home	170
6.5	Like and Unlike Charges	171
6.6	Electric Charges and Magnetic Poles	174
6.7	Polarization	177
6.8	Charging by Induction	179
6.9	Coulomb's Law	179
6.10	Electrostatic Interaction and Newton's Third Law	182
6.11	Sharing Charge Between Two Spheres	183

6.12	Conservation of Charge	184
6.13	Electrical Field Strength	185
6.14	Superposition	186

CHAPTER 7 CURRENT ELECTRICITY 188

7.1	Introduction	188
7.2	Static or Current Electricity First?	189
7.3	Current Electricity as Charge in Motion	190
7.4	Formation of Basic Circuit Concepts	194
7.5	Phenomenology of Simple Circuits	198
7.6	Historical Development of Ohm's Law	200
7.7	Teaching Electrical Resistance and Ohm's Law	204
7.8	Current: A Bulk or Surface Phenomenon?	205
7.9	Building the Current-Circuit Model	206
7.10	Conventional Current Versus Electron Current	208
7.11	Not Every Load Obeys Ohm's Law	209
7.12	Free Electrons in Metals	210
7.13	Sample Homework and Test Questions	214

CHAPTER 8 ELECTROMAGNETISM 218

8.1	Introduction	218
8.2	Oersted's Experiment	219
8.3	Forces Between Magnets and Current Carrying Conductors	222
8.4	Ampere's Experiment	223
8.5	Mnemonics and the Computer	225
8.6	Faraday's Law in a Multiply Connected Region	226
8.7	Faraday's Criticism of Action at a Distance	227
8.8	Infancy of the "Field" Concept	230
8.9	Laboratory Measurement of a Value of B	233

CHAPTER 9 WAVES AND LIGHT 234

9.1	Introduction	234
9.2	Particle and Propagation Velocities	234
9.3	Graphs	235
9.4	Transverse and Longitudinal Pulse Shapes	237
9.5	Reflection of Pulses	238
9.6	Derivation of Propagation Velocities	241
9.7	Velocity of Propagation of a Kink on a String	242
9.8	Propagation Velocity of a Pulse in a Fluid	244
9.9	Surface Waves in Shallow Water	247
9.10	Transient Wave Effects	250

9.11	Wave Fronts and Rays in two Dimensions	251
9.12	Periodic and Sinusoidal Wave Trains	252
9.13	Two-Source Interference Patterns	253
9.14	Two-Source Versus Grating Patterns	254
9.15	Young's Elucidation of the Dark Center in Newton's Rings	256
9.16	Specular Versus Diffuse Reflection	257
9.17	Images and Image Formation: Plane Mirrors	258
9.18	Images with Thin Converging Lenses	260
9.19	Novice Conceptions of the Nature of Light	263
9.20	Phenomenological Questions and Problems	263

CHAPTER 10 EARLY MODERN PHYSICS 265

10.1	Introduction	265
10.2	Historical Preliminaries	266
10.3	Prelude to Thomson's Research	271
10.4	Thomson's Experiments	272
10.5	Thomson's Inferences	275
10.6	Homework on the Thomson Experiment	277
10.7	The Corpuscle of Electrical Charge	278
10.8	From Thomson's Electron to the Bohr Atom	279
10.9	Photo-Emission and the Photon Concept	285
10.10	Einstein's Paper on the Photon Concept	290
10.11	Bohr's Model of Atomic Hydrogen	292
10.12	Introducing Special Relativity	301
10.13	Written Homework on the Thomson Experiment	308
10.14	Written Homework on the Bohr Atom	313

CHAPTER 11 MISCELLANEOUS TOPICS 318

11.1	Introducing Kinetic Theory	318
11.2	Assumptions of Kinetic Theory	320
11.3	Hydroststic Pressure	327
11.4	Visualizing Thermal Expansion	329
11.5	Estimating	329
11.6	Significant Figures	330
11.7	Precision, Accuracy, and Significant Differences	331
11.8	Distribution Functions	332
11.9	Guidance in Introductory Laboratory	333
11.10	Cultivating Insight and Inquiry in Laboratory	335
11.11	Mathematical Physics for Gifted Students	339
11.12	Chaos	342

CHAPTER 12	ACHIEVING WIDER SCIENTIFIC LITERACY	344
12.1	Introduction	344
12.2	Marks of Scientific Literacy	345
12.3	Operative Knowledge	347
12.4	General Education Science Courses	349
12.5	Illustrating the Nature of Scientific Thought	352
12.6	Connections to Intellectual History	358
12.7	Variations on the Theme	361
12.8	Aspects of Implementation	362
12.9	The Problem of Cognitive Development	365
12.10	The Problem of Teacher Education	365
12.11	A Role for the Computer	369
12.12	Learning from Past Experience	370
CHAPTER 13	CRITICAL THINKING	375
13.1	Introduction	375
13.2	A List of Processes	376
13.3	Why Bother with Critical Thinking?	382
13.4	Existing Level of Capacity for Abstract Reasoning	384
13.5	Can Capacity for Abstract Reasoning Be Enhanced?	385
13.6	Consequences of Mismatch	387
13.7	Ascertaining Student Difficulties	389
13.8	Testing	390
13.9	Some Thoughts on Faculty Development	390
	BIBLIOGRAPHY	393
	INDEX TO PART I	405

Chapter 1

Underpinnings

1.1 INTRODUCTION

Several fundamental gaps in the background of students may seriously impede their grasp of the concepts and lines of reasoning that we seek to cultivate from the beginning of an introductory physics course. These gaps, having to do with understanding the concepts of “area” and “volume” and with reasoning involving ratios and division, are often encountered, even among students at the engineering physics level.

In principle, these gaps should not exist because the ideas are dealt with, and should have been mastered, at earlier levels in the schools. It is an empirical fact, however, that such mastery has not been achieved, and ignoring the impediment is counterproductive.

Unfortunately, it is illusory to expect to remediate these difficulties with a few quick exercises, in artificial context, at the start of a course. Most students can be helped to close the gaps, but this requires *repeated* exercises that are spread out over time and are integrated with the subject matter of the course itself. This statement is *not* a matter of conjecture; it reflects empirical experience our physics education research group at the University of Washington has encountered repeatedly [Arons (1976), (1983b), (1984c)].

This chapter describes some of the learning difficulties that are involved in the development of a number of underpinnings, including arithmetical reasoning, and suggests exercises that can be made part of the course work.

1.2 AREA

The concept of area underlies the formation of many basic physical concepts, such as pressure, stress, energy flux, and coefficients of diffusion and heat conduction. It underpins all the ratio reasoning associated with geometrical scaling. Furthermore, it is essential to the interpretation of velocity change as area under the graph of acceleration versus clock reading, to the interpretation

of position change as area under the graph of velocity versus clock reading, to the definitions of work and impulse, and to the interpretation of integrals in general.

If you ask students how one arrives at numerical values for “area” or “extent of surface,” many—if they have any response at all—will say “length times width.” If you then sketch some very irregular figure without definable length or width and ask about assigning a numerical value to the area of the figure, very little response of any kind is forthcoming. Students who respond in this way have not formed a clear operational definition of “area.”

The reason for this is fairly simple: Although the grade school arithmetic books, when they introduce the area concept, do have a paragraph about selecting a unit square, imposing a grid on the figure in question, and counting the squares within the figure, virtually none of the students have ever gone through such a procedure themselves in homework exercises. They have never been asked to define “area.” All they have ever done is deal with the end results, calculating areas of regular figures such as squares, rectangles, parallelograms, or triangles, using memorized formulas that they no longer connect with the operation of counting the unit squares, even though this connection may have been originally asserted. They are unable to account for the origin of the formulas they are invoking.

Furthermore, virtually none of the students have had any significant exposure to the notion of operational definition. They have had little or no practice in defining a term by reference to shared experience or by describing, in simple words of prior definition, the actions through which one goes to develop the numerical value being referred to in the name of a technical concept.

1.3 EXERCISES WITH “AREA”

In introductory physics teaching, it is desirable to invoke the area concept at the earliest possible opportunity. Students should be led to articulate the operational definition in their own words—and to do so on tests. (This is an excellent opportunity to introduce the concept of operational definition in a context that is familiar and relatively unthreatening.) The fact that they had been using the technical term “area” without adequate mastery of the concept behind it makes a salutary impression on many students.

Homework and test problems should give students opportunity to execute the operations they describe in the definition, right through the selection of the unit square, superposition of the grid on the figure in question, and actually counting the squares. The operation of counting must involve the estimation of squares contained around the periphery of the figure. To many students the necessity of estimating the fractions appears in some sense “sinful,” since it involves “error” and is not “exact,” as seems to be the value obtained from a formula. The actual experience of counting and estimating should begin with “pure” areas, that is, surface extent of arbitrarily and irregularly shaped

geometrical figures. Then, as soon as it becomes appropriate, the exercises should be extended to measurement and interpretation of areas under v versus t and under a versus t graphs. (This, of course, adds the arithmetical reasoning associated with the dimensionality of the coordinates.)

In calculus-physics courses, the latter exercises should be explicitly linked with the mathematical concept of “integral.” Although this might seem so obvious as to be not worthy of mention, many students have not actually established this connection even though they may be taking, or may have completed, a calculus course. Although they have been *told*, perhaps many times, that the integral can be interpreted as an area, the idea has not registered because it has not been made part of the individual student’s concrete experience, and they have never had the opportunity to articulate the idea in their own words.

Such exercises should be repeated still later when the context begins to involve “work” and “impulse.” It is only such recycling of ideas over fairly extended periods of time, reencountered in increasingly rich context, that leads to a firm assimilation in many students.

In algebra-based physics courses, the concept of “integral” is not at hand and is not necessary. Dealing with the areas, however, breaks the shackles to eternally constant quantities and shows the students how physics can easily and legitimately deal with continuous change. “Capturing the fleeting instant” was one of the great intellectual triumphs of the seventeenth century, and students can be given some sense of this part of their intellectual heritage through calculations that they can easily make without the necessity of a formal course in the calculus.

1.4 VOLUME

Initially, most students have the same difficulty with “volume” as with “area.” They grasp for formulas without having registered an operational definition of the concept. As a result, quite a few students do not, in fact, discriminate between area and volume; they use the words carelessly and interchangeably as metaphors for size.

Once the operational definition of “area” has been carefully developed and anchored in the concrete experience of counting squares, however, the operational definition of “volume” can be elicited relatively easily. The analogy to “area” is readily perceived, and the counting of unit cubes is quickly accepted.

1.5 MASTERY OF CONCEPTS

It should be emphasized that mastery of the operational definitions of “area” and “volume” up to the point of recognizing the counting of unit squares or cubes is only a beginning; it is still far short of the ability to use the

concepts in more extended context. At this stage, for example, some students (particularly those who have had little or no prior work in science) do not discriminate between mass and volume.¹ Many students, including those in engineering-physics courses, are, at this stage, still unable to compare final with initial areas or volumes when the linear dimensions of an object have been scaled up or down.

The problem of scaling is a particularly important one. It involves ratio reasoning and will be discussed in more detail in Section 1.12.

1.6 RATIOS AND DIVISION

One of the most severe and widely prevalent gaps in cognitive development of students at secondary and early college levels is the failure to have mastered reasoning involving ratios. The poor performance reproducibly observed on Piagetian tasks of ratio reasoning has become well known since the early 1970s [McKinnon and Renner (1971); Karplus, et al. (1979); Arons and Karplus (1976); Chiappetta (1976)]. This disability, among the very large number of students who suffer from it, is one of the most serious impediments to their study of science.

For convenience, I separate reasoning with ratios and division into two levels or stages: (1) verbally interpreting the result obtained when one number is divided by another; (2) using the preceding interpretation to calculate some other quantity.

1.7 VERBAL INTERPRETATION OF RATIOS

Reasoning with ratios and division requires, as a first step, the capacity to interpret verbally the meaning of a number obtained from a particular ratio. The verbal interpretations are somewhat different in different contexts. Many students are deficient in this capacity and need practice in interpreting ratios in their own words.

In the primitive case in which the numbers have not been given specific physical meaning, we interpret the result of, say, $465/23$, as the number of times 23 is contained in 465. This may sound like a trivial statement, but it is not. Most students have memorized (successfully or unsuccessfully, as the case may be) the algorithm of division but have never been given the opportunity to recognize it as a shorthand procedure for counting successive subtractions of 23 from 465. Thus they do not see the operation of division in perspective or translate it into simpler prior experience. The phrase “goes into” is memorized without relation to other contexts. Those who have not developed this

¹For evidence concerning this assertion and for strategies that help students achieve such discrimination see McDermott, Pitenick, and Rosenquist (1980); McDermott (1980); McDermott, Rosenquist, and van Zee (1983).

perspective should be given the opportunity to count the successive subtractions and to begin to see what they are doing in the memorized algorithm. They should finally have to tell the whole story in their own words. Students can be led to perform such inquiry by means of the hand calculator even if the ancient pencil-and-paper modes have vanished into mists of unfamiliarity.

At a next higher level of sophistication, we may be dealing with a ratio of dimensionally identical quantities, for example, L_2/L_1 , the ratio, say, of the heights of two buildings, or of distances from a fulcrum in balancing, or the linear scaling of a geometrical figure. Here the numerical value of the ratio serves as a *comparison*: it tells us how many times larger (or smaller) one length is compared to the other.

Next we encounter division of dimensionally *inhomogeneous* quantities: mass in grams divided by volume in cubic centimeters; position change in meters divided by a time interval in seconds; dollars paid divided by number of pounds purchased. Here the result of division tells us how much of the numerator is associated with one unit of whatever is represented in the denominator.

Finally, if we have 500 g of a material that has 3.0 g in each cubic centimeter, the numerical value of $500/3.0$ tells us how many “packages” of size 3.0 g are contained in the 500 g sample. Since each such “package” corresponds to one cubic centimeter, we have obtained the number of cubic centimeters in the sample.

1.8 EXERCISES IN VERBAL INTERPRETATION

Many students have great difficulty giving verbal interpretations such as those illustrated in the preceding section since they have almost never been asked to do so. Without such practice in at least several different contexts, many students do not think about the meaning of the calculations they are expected to carry out, and they take refuge in memorizing patterns and procedures of calculation—manipulating formulas, rather than penetrating to an understanding of the reasoning. As a consequence, when they find themselves outside the memorized situations, they are unable to solve problems that involve successive steps of arithmetical reasoning.

Explaining or telling students who are in such difficulty the meaning of particular ratios, however frequently or lucidly this may be done, has very little effect. It is necessary to ask questions that lead the students to articulate the interpretations and explanations in their own words. In the paragraphs that follow are some typical excerpts from such conversations.

Suppose students having difficulty with a problem involving the use of the density concept are asked: “We took the measured mass (340 g) of an object and divided it by the volume (120 cm^3). How do you interpret the number $340/120$? Tell what it means, using the simplest possible words.” Some will answer “That is the density.” These students have not separated the technical

term, the *name* of the resulting number, from the verbal interpretation of its meaning. (This involves an important cognitive process that will be discussed in another chapter.)

When it is pointed out that the name is not an interpretation, some students will say “mass per volume”; others might say “the number of grams in 120 cubic centimeters.” (Exactly parallel statements are likely to be given if the ratio is position change divided by time interval.) Very few students having trouble with the original problem will give a simple statement to the effect that we have obtained the number of grams in *one* cubic centimeter of the material.

One can now adopt the strategy of going back to some more familiar context: “Suppose we go to a store and find a box costing \$5.00 and containing 3 kg of material. What is the meaning of the number $5.00/3$?” Some students will still say “That is how much you pay for 3 kg” but, in this more familiar context, many will recognize that we have calculated how many dollars we pay for one kilogram. (The former group is in need of further dialog, using more concrete examples, before a correct response is produced.) One can now try to get the students to the generalization that in such situations the resulting number tells us “how many of these (in the numerator) are associated with *one* of those (in the denominator).”

If one then asks: “In the case of the box costing \$5.00 and containing 3 kg, suppose we now consider the number $3/5.00$. In light of what we concluded in the previous example, does this number have an interpretation?” Many students, including some who gave the correct interpretation of $5.00/3$, now encounter difficulty. Some revert to earlier locutions such as “how many kilograms you get for \$5.00”; many consider the number meaningless or uninterpretable.

In such instances there seem to be two difficulties superposed: (1) although the students may have previously been given some opportunity to think about or calculate “unit cost” (how much we pay for one kilogram), they rarely, if ever, have been asked about the inverse (how much one gets for *one* dollar). (2) $5.00/3$ involved the division of a larger number by a smaller one. To many students this is more intelligible and less frightening than the fraction $3/5.00$.

After students have been led through the parallel interpretation of both ratios, one can usually go back to a case such as mass divided by volume or change of velocity divided by time interval and elicit a correct interpretation of the new ratio and its inverse. Then one can elicit the generalization being sought, namely, that such a ratio tells us how much of the numerator is associated with *one* unit of whatever is represented in the denominator. It is essential, however, to elicit the word “one”; use of the word “per” by the student is no assurance that he or she understands the concept (see the discussion in the next section).

1.9 COMMENT ON VERBAL EXERCISES

Note the strategy being employed in the dialogs suggested in the preceding section: although some students have responded previously to problems such as “calculate the cost of one kilogram if 3 kg cost \$5.00,” very few students have ever been confronted with the ratio and asked to interpret it in words, that is, they have never reversed the line of thought, traversing it in the direction opposite to that previously experienced.

In Piagetian terminology, the term “operations” denotes reasoning processes that can be reversed by the user. Thus students who can calculate the unit cost but do not recognize the interpretation of the ratio are not reversing the reasoning and have not brought it to the “operations” level. Leading them to reverse the direction of reasoning turns out to be a useful tool for helping them master the reasoning. (This idea will be discussed in more general terms in a subsequent chapter.)

Complete control of the interpretation of ratios is rarely attained with just one short sequence of exposure as outlined above. Many students must have the experience of carrying through the same kind of reversible reasoning in several additional contexts (e.g., what is the meaning of the number obtained in dividing the circumference of a circle by its diameter? If 16 g of oxygen combine with 12 g of carbon, what is the meaning of $16/12$? Of $12/16$? If a laboratory cart travels 180 cm in 2.3 s, what is the meaning of $180/2.3$? Of $2.3/180$? etc.) before they fully assimilate it.

A word of warning: If a teacher accepts casual use of the word “per”—particularly the incorrect and meaningless “mass per volume,” which was quoted in the preceding section—he or she falls into a trap. Even though it contains only three letters, “per” is a technical term, and very few of those students who are having trouble with arithmetical reasoning know what it means. They inject it into a response only because they have a vague memory that “per” frequently turns up for some obscure reason in division, but they do not explicitly translate it into simpler words such as “in,” “for each,” “corresponds to,” “goes with,” “combines with,” “is associated with.”

Even if students correctly say “mass per *unit* volume” rather than “mass per volume” in interpreting M/V , there is no conclusive assurance that they really understand the meaning. Some do, but others have merely memorized the locution.² It is important to lead all students into giving simple interpre-

²Tobias (1988) notes a similar problem, stemming from inattention on the part of teachers, in connection with the word “of”:

A number of [students] reported getting lost during lessons on multiplication and division of fractions, and as they talked about this, I began to notice an ambiguity in use of the word “of.” They had been instructed that the word “of” in expressions like one-third of three-quarters always means multiply. But this, they remembered, felt wrong or confusing. “Of” felt more like division. Indeed they were right. Words connote as well as denote. The word “of,” in fact, means multiply only in one narrow context within mathematics.

tation in everyday language before accepting a regular use of “per.” Many students do not know what the word “ratio” means. Those having difficulty with reasoning and interpretation should always be asked, at an early stage, for the meaning of the word if they, the text, or the teacher invoke it.

It is also worth noting that the interpretations of division being illustrated underlie many of the manipulations of elementary algebra and are particularly relevant to the translation of verbal problems into the corresponding algebraic equations and vice versa. Remediating student difficulties with verbal interpretation of ratios eventually enhances students’ ability to use elementary algebra.

1.10 ARITHMETICAL REASONING INVOLVING DIVISION

Verbal interpretations like those illustrated in the preceding section (how much of the numerator is associated with *one* unit of whatever is represented in the denominator) are only the first step in a sequence and involve only one of the several interpretations of the meaning of a result of dividing one number by another. The next fruitful step is made through such questions as: “We have 800 g of material having a density of 2.3 g/cm³. What must be the volume occupied by the sample?”

The first impulse of many students is to manipulate the density formula $\rho = M/V$. (In fact, if the word “density” is not used in the statement of the question and one merely says that the material has 2.3 g in each cubic centimeter, quite a few students are completely lost, not knowing what to do when they have not been cued as to a formula.) An investigation of what is happening in manipulation of the formula reveals what Piaget would characterize as an essentially “concrete operational” response. In many instances, the students are not reasoning either arithmetically or algebraically but are simply rearranging the symbols, as though they were concrete objects, in patterns that have become familiar. Obtaining a correct answer to the initial question does not necessarily indicate a grasp of the attendant arithmetical reasoning.³

Students should be led to articulate something like the following story: What does 2.3 g/cm³ *mean*? The quantity 2.3 is the number of grams in *one* cubic centimeter. We can think of 2.3 g as a clump or package. If we find how many such packages there are in 800 g of the material, we obtain the total

³It should be pointed out that classical “proportional reasoning” (e.g., object A has a height of 8 measured in units of length of a small paper clip. Object B has a height of 12 in the same units. Object A has a height of 6 measured in units of length of a larger paper clip. What would be the height of B measured in large paper clips?) suffers from similar problems. Many students memorize the “this-is-to-this-as-that-is-to-that” routine and manipulate the given numbers as concrete objects in a spatial arrangement, frequently doing so incorrectly. Again, a correct result is not firm evidence of understanding the line of reasoning.

number of cubic centimeters in 800 g because each package corresponds to one cubic centimeter.

Similarly, when asked to find the diameter of a circle having a circumference of 28 cm, students should be led to argue that, since each “package” of 3.14 cm in the circumference corresponds to one centimeter in the diameter, we must find how many packages of size 3.14 are contained in 28. Manipulation of the formula $C = \pi D$, however correctly, does not testify to understanding of the meaning of π or to grasp of the underlying arithmetical reasoning.

One such exposure does not usually provide full remediation to students who have this difficulty. Repetition is essential, but repetition without some alteration of the context simply encourages memorization. One way of altering the context sufficiently to make the repetition nontrivial is as follows: “We have a block consisting of 5000 g of material having a density of 2.3 g/cm³. Suppose we add 800 g of the same material to the block. By how much have we increased the volume of the block?” (Similarly, one alters the circle problem by adding 28 cm to the circumference of a circle having some arbitrary initial diameter, large or small, and asking for the increase in diameter.)

Many students initially see these problems as entirely different from the original versions. They painstakingly calculate, for example, the volume of a 5800 g block and subtract the volume of a 5000 g block. When they are led to realize that $800/2.3$ gives the answer to both versions, they make a significant stride toward mastery of the underlying reasoning, especially when they additionally recognize that the circle problems are exactly the same as the density problems.

To summarize: linguistic elements play an essential, underlying role in the development of the capacity for arithmetical reasoning with ratios and proportion. This observation is explicitly supported by Lawson, Lawson, and Lawson (1984) who remark that “a necessary . . . condition for the acquisition of proportional reasoning during adolescence is the prior internalization of key linguistic elements or argumentation.” Failure to provide this linguistic experience in the schools underlies much of the difficulty students experience, and much of the “fear of mathematics” that we observe, at high school and college levels. The pace at which verbal security can be conveyed at the latter levels is no greater than the pace required at earlier age. This problem will not be rectified until we, in the colleges and universities, produce elementary school teachers who have mastered arithmetical reasoning sufficiently thoroughly to lead their pupils into articulating lines of reasoning and explanation in their own words. This is not currently being achieved in sufficiently large measure.

1.11 GRAPHS AND ARITHMETICAL REASONING

A powerful way of helping students master a mode of reasoning is to allow them to view the same reasoning from more than one perspective. In the case of arithmetical reasoning, a very useful alternative perspective is that

of graphical representation. Consider, for example, the different situations illustrated in the graphs of Fig. 1.11.1.

Students should not be confronted with these graphs all at once in some remedial orgy. They should be led into building up the representations in homework problems whenever the situations arise in the normal sequence of the course work. This allows for spiralling back to the modes of thinking and spreads the encounters out over weeks of time; both the spiralling back and the time spread are essential for effective assimilation. In each encounter, they should have to interpret the representations in their own words. For example:

1. In Fig. 1.11.1(a) each line represents a different substance; the steepness (or slope) of the line is the number M/V and is interpreted as the number of grams in one cubic centimeter if the units are grams and cubic centimeters, respectively; in any straight-line relationship the amount added along the vertical axis is always the same for equal steps along the horizontal axis; when the graph is not a straight line, the steps along the vertical axis are not equal under such circumstances.

2. The steepness of such straight lines is frequently a *property* of the object or system being described. In Fig. 1.11.1(a) the property is called “density of the substance”; in (b) it is called “concentration of the solution”; in (c) it is called “inertial mass of the object”; in Fig. 1.11.1(d) it is called “coefficient of friction between the two surfaces,” and so on.

3. In most of the graphs, different systems or objects possess their own different numerical values of the property in question, and there are different straight lines for different objects. Figure 1.11.1(g), however, illustrates the remarkable fact that the steepness 3.14, to which we give the name π , is a property that *all* circles have in common, and there is only one straight line!

4. The problems in Section 1.10 that involve arithmetical reasoning with the concept of density and π can be represented and interpreted on Figs. 1.11.1(a) and 1.11.1(g), and students should be led to do so. In order to calculate the total volume of a sample of known mass and known density, or the diameter of a circle of known circumference, one can use the straight line from the origin to the mass or circumference in question. Calculations of the volume *added* to a sample, or the *increase* in diameter of a specified circle, are represented by the small dashed triangles in the respective figures. The graphical representation helps reinforce the insight that a given change along the horizontal axis produces a corresponding, fixed change along the vertical axis regardless of whether the shift is started at the origin or elsewhere along the line.

It helps to dramatize this idea by asking students to imagine a string around the equator of the earth, forming a circle with a circumference of 40,070 km. Now suppose we increase the length of the string by 6.0 m; what will be the *increase* in the diameter of the circle it forms? What would be the

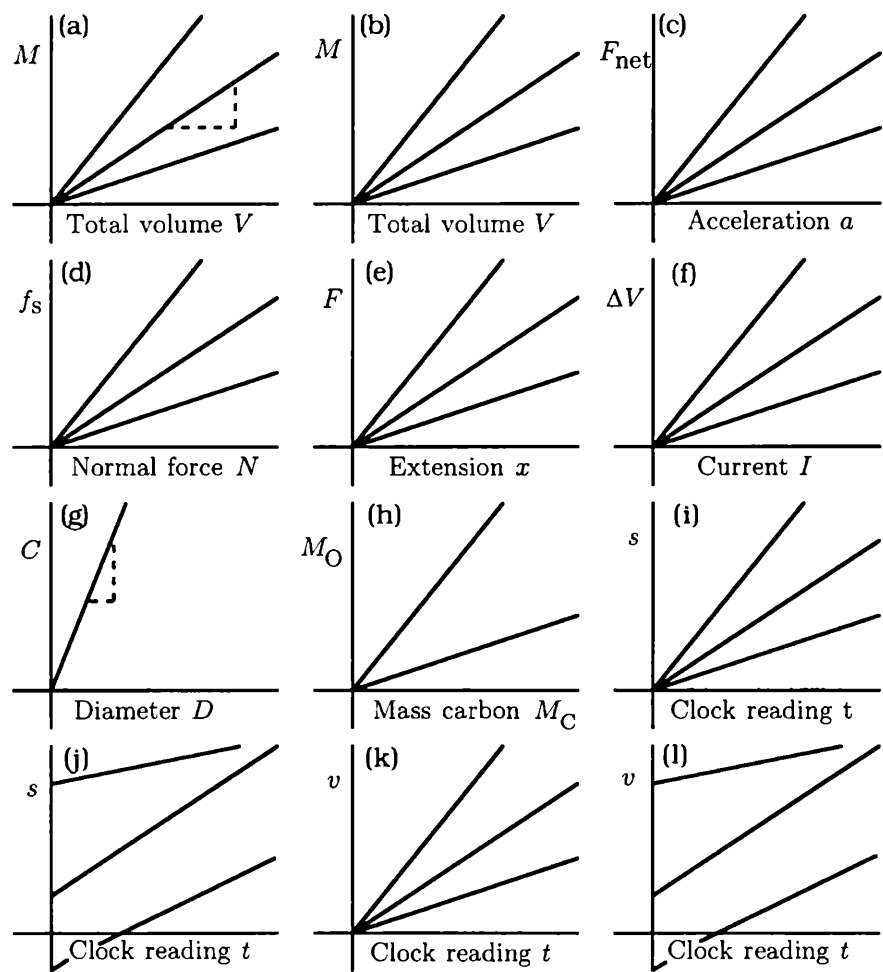


Figure 1.11.1 Linear relations and arithmetical reasoning. (a) Total mass M versus total volume V of three different homogeneous substances. (b) Total mass M of solute versus total volume V of solvent for three different solution concentrations using the same solute and the same solvent. (c) Net force F_{net} versus acceleration a for three different objects (having different inertial masses) in rectilinear motion. (d) Maximum static frictional force f_s versus normal force N for sliding involving three different pairs of surfaces. (e) Applied force F versus resulting extension x from relaxed condition for three different springs obeying Hooke's law. (f) Potential difference ΔV versus current I for three different electrical conductors obeying Ohm's law. (g) Circumference C versus diameter D for *all* circles. (h) Total mass of oxygen M_O versus total mass of carbon M_C in samples of carbon dioxide and carbon monoxide. (i) Rectilinear motion: position s versus clock reading t for three different objects all having position $s = 0$ at $t = 0$. (j) Rectilinear motion: position s versus clock reading t for three different objects all having different values of s at $t = 0$. (k) Rectilinear motion: instantaneous velocity v versus clock reading t for objects having zero velocity at $t = 0$. (l) Rectilinear motion: instantaneous velocity v versus clock reading t for objects having different velocities at $t = 0$.

increase in diameter if we added 6.0 m to the circumference of a circle having an initial circumference of 8.0 cm?

5. In addition to providing further exercises with parallel arithmetical reasoning in entirely different context, Figs. 1.11.1(i) and 1.11.1(k), on the one hand, juxtaposed against Figs. 1.11.1(j) and 1.11.1(l), on the other, illustrate the difference between a direct proportion and a linear relation that is *not* a direct proportion. Very few students have formed this distinction explicitly, and many texts and teachers confuse the issue by careless use of the terminology.

Combining the modes of reasoning described in Section 1.10 with the parallel graphical representations described in this section, pointing out the connections explicitly, and requiring the students to describe them in their own words strongly serve to enhance and secure students' grasp of both reasoning with division and the interpretation of straight-line graphs. One might even say that the superposition of the two perspectives is nonlinear.

1.12 SCALING AND RATIO REASONING

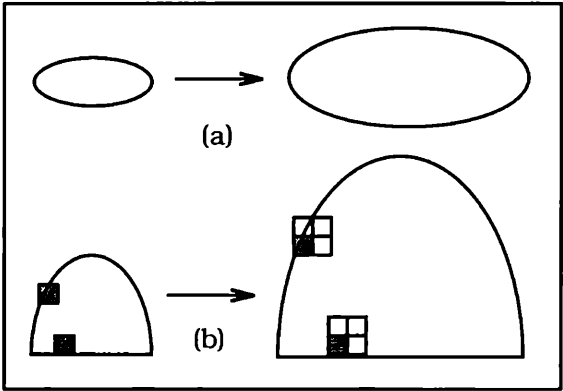
Suppose we double the linear dimensions of, say, a statue: What will happen to the circumference of an arm? To the cross-sectional area of a leg? To the total surface area? To the volume of the required mold for casting? The great majority of students, including those in engineering-physics courses, have very serious difficulty with such questions, and the difficulty is compounded if the scale factor has a noninteger value. Many will guess, without thinking or analysis, that areas and volumes will increase by the given linear factor. They find themselves helpless in confronting the scale ratio alone without the actual initial dimensions of the object. They have no idea what to do in the absence of formulas for the relevant areas and volumes.

There are two principal difficulties behind this deficiency. The first has been discussed in Sections 1.2 and 1.3 above: the fact that the students have not been helped to form explicit operational definitions of "area" and "volume." The second difficulty resides in the fact that very few students have formed any conception of the basic *functional* relation between area and linear dimensions, on the one hand, or between volume and linear dimensions, on the other. Memorizing and using formulas for regular figures does *not* help form this conception. Hence, students are unaware that all areas vary as the square of the length scale factor, and that volumes vary as the cube, regardless of regularity or irregularity of shape and regardless of existence or nonexistence of a formula.

Even if they are vaguely aware of the functional relations, they are unable to deal with them in terms of ratios, that is, they do not think in terms of what mature scientists and engineers call "scaling." Remediation must come by first filling the gaps outlined in the preceding sections. Then students can be led to visualize what happens to unit squares as the dimensions of an

arbitrary plane figure are doubled as illustrated, for example, in Fig. 1.12.1: any one unit square in the smaller figure expands into four such squares in the larger, whether in the interior of the figure or along the periphery. The reverse takes place when scaling down rather than up. Students should then sketch for themselves what happens when the scale factor is 3 or 4 rather than 2.

Figure 1.12.1 Two different plane figures scaled up by a factor of 2 in linear dimensions. In (b) it is shown that any one unit square in the smaller figure expands into *four* such squares in the larger figure and that this happens throughout the entire figure, including the periphery.



Those students, and there are many, who have difficulty extending the idea to noninteger scale factors should be led to sketch Fig. 1.12.2 in which dimensions are increased by a factor of 1.5, and one can readily confirm, by actually counting the squares, that the area increases by the factor $(1.5)^2/1$ since there are 2.25 unit squares in the larger figure.

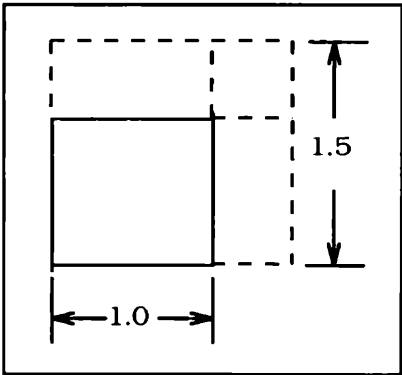


Figure 1.12.2 When the linear dimensions of a square are scaled up by a factor of 1.5, the new square contains 2.25 original squares.

Then one must extend the thinking to three dimensions and lead students to generalize the cubic functional relation for volume. Exercises can then be given in which areas and volumes are scaled up or down, as well as exercises in which the reasoning must be reversed, that is, given the ratio by which area has been scaled up or down, what are the corresponding scale factors for length and volume? The great majority of students initially have very severe difficulty with the latter question; the necessity of taking roots instead of raising to powers turns out to be a formidable obstacle.

If these exercises, however, are confined to an initial short remedial period and are stated exclusively in terms of the abstractions “area” and “volume” without connection to visualization of concrete objects, without review of operational definition, and without being embedded in richer context, very little learning takes place; the calculational procedures are temporarily memorized and are quickly forgotten.

It is important to return from time to time to scaling in different substantive contexts, giving the students the chance to encounter a variety of applications: the role of surface-to-volume ratio in determining rate of solution or in comparing metabolic rates in cells or in large and small animals; the fact that the leg bones of elephants must have a disproportionately larger diameter than do those of horses in order to sustain the increased weight; what happens to the density of gas in a balloon if the linear dimensions of the balloon are doubled without addition or escape of gas?

Then, as more physics subject matter is developed, such thinking in terms of ratios should be extended to other and more abstract functional relations:

We have a bob on a string in horizontal circular motion. What happens to the centripetal force acting on the bob if the angular velocity is increased by a factor of 1.6, other quantities being held constant? What must be done to the tangential velocity in order to decrease the centripetal force by a factor of 2? What happens to the centripetal force if the mass of the bob and the radius of the circle are both tripled without change in angular velocity? If the tangential velocity is doubled, what must be done to the radius to keep the centripetal force unchanged?

If the magnitude of the force acting on a certain lever arm is decreased by a factor of 2.3, what must happen to the length of the lever arm to keep the torque unchanged?

If, in an interaction between point charges, one charge is increased by a factor of 3.5, what must be done to the separation between the charges to keep the force of interaction unchanged?

In all these examples, students initially exhibit *very* strong resistance to doing the thinking in terms of ratios and functional relationships. They want initial numerical values, and they want to substitute into the formulas without having to think through the ratios and without having to decide whether the quantity in question is going to increase or decrease. The resistance can be overcome only through repeated exposure and practice. It is well known to most college teachers that upper division engineering students and science majors are very deficient in ability to estimate and to do ratio reasoning of the kind described above. The reason for this deficiency is very simple: the students have been given little or no practice in such thinking, and the capacity does not develop spontaneously. When the breakthrough is attained, however,

after repeated encounter, the self-confidence and self-respect of the students increase immeasurably, and their rate of progress is clearly enhanced.

Some teachers may remember the beautifully written Part I of the first two editions of the high school course *PSSC Physics*, with its fine overview of the science to be developed in more detail in the subsequent parts. Part I, which happened to be deeply infused with scaling and ratio reasoning, was deemed a “failure” and was removed in subsequent editions in the belief that the overview was premature and too sophisticated. I originally shared this view but, in retrospect, I have come to believe that the problem with Part I was not so much in its subject matter as in the fact that neither the students nor the high school teachers were ready for the ratio reasoning, scaling, and estimating that permeated the sequence. The content was obscured by the impenetrability of the ratio reasoning.

If we do not help our students penetrate this obstacle, we shall never get them to the point of willingness to estimate or to make order of magnitude analyses and predictions, since such reasoning usually involves ratios, scaling, and functional relation. One hears frequent complaints that even physics majors and graduate students are gravely deficient in these skills. They are indeed deficient in this respect, and the reason is that they have had virtually no practice. (See Section 11.5 for references to papers giving problems and exercises on estimating.)

1.13 ELEMENTARY TRIGONOMETRY

Although in the more mathematically sophisticated sense sine, cosine, and tangent of an angle are to be regarded as functions, the students first encounter and use them as simple ratios of lengths of sides in right triangles. They laboriously memorize the standard definitions and use them as formulas to be rearranged by algebraic manipulation whenever a calculation on a right triangle is to be made.

The functional generalization is not necessary at this juncture and is not likely to be helpful. Students should first be led to see sines and cosines as simple *fractions*. If one multiplies the hypotenuse by the fractions, one obtains the lengths of the sides opposite and adjacent to the angle, respectively. This broadens the perspective by giving the students an alternative view of what the names “sine” and “cosine” stand for; it helps them think directly and concretely about the lengths—thinking that they are not doing when they mechanically and abstractly rearrange the standard formulas. The broadened perspective, however, rarely arises spontaneously; it must be deliberately induced by the teacher. This is clearly a matter of drill and practice that could readily be delivered via microcomputer.

Instructors should be explicitly aware of another basic aspect of trigonometry in which students are markedly deficient, even if they have had exposure in high school, namely that of radian measure. They have rarely, if ever, used

radian measure in any significant context. They may have temporarily memorized a definition and used it in trivial conversion exercises, but they have not been shown *why* this dimensionless angular measure is useful, important, and even necessary. This deficiency is best remedied not by launching into a “review” at the beginning of a course but by showing the need for radian measure when an appropriate context is encountered. Hence, the approach to radian measure will be discussed in more detail in Chapter 4 on two-dimensional motion.

1.14 HORIZONTAL, VERTICAL, NORTH, SOUTH, NOON, MIDNIGHT

Very few students can give intelligible operational definitions of the terms appearing in the title of this section. If one asks students, “What is meant by the term ‘vertical’? How would you proceed to establish the vertical direction right here in this place?”, a frequently occurring response is, “Perpendicular to the ground.” If one then suggests going over to the steep slope nearby and erecting a perpendicular to the ground, the student recognizes an inconsistency but rarely sees any way out. It takes some minutes of hinting and questioning to draw out a proposal to hang a weight on a string and make a plumb bob. Relatively few students in this day and age have heard the term “plumb bob” or know what it means; nor do they know the meaning of the word “plumb” by itself.

Another acceptable, albeit more cumbersome, approach would be to establish the horizontal by means of a carpenter’s level and then erect the perpendicular, but this suggestion very rarely emerges.

Similar discussions need to be conducted with respect to the other terms cited above. If asked how the local north direction is defined and established, most students refer to the magnetic compass as though this were a primary definition. They do not connect “north” with either the direction of the celestial pole or the shortest shadow cast by a vertical stick.

If asked about the meaning of “local noon,” most students are likely to refer to the sun being “directly overhead” without awareness that in the latitude at which most of them live the sun never passes through the zenith. When they are led to realize that the sun does not pass through the zenith, they can be led to the shadow of the vertical stick as a simple device for determining highest elevation of the sun and thus to the definition of local noon. A fruitful discussion question then resides in “What significance, if any, do you see in the fact that the directions of the North Star and the shortest shadow of the vertical stick coincide? Is this simply an accident or might it have deeper meaning?”

Again, such discussions are ineffective in an a priori review. They register most effectively if the student is challenged on the meaning of each term when

it first arises in some specific context of problem or reading or discourse. The terms are so familiar and frequently invoked that the student has lost all sense of the fact that he or she does not really know what they mean. The necessity of groping for a simple operational definition of such familiar terms is, at first, embarrassing but provides a very salutary intellectual experience.

1.15 INTERPRETATION OF SIMPLE ALGEBRAIC STATEMENTS

Lochhead and Clement and their co-workers at the University of Massachusetts, Amherst, have studied the difficulties many individuals have with the translation of simple algebraic statements from words to an equation and from an equation to words [see Clement, Lochhead, and Monk (1981); Rosnick and Clement (1980)]. A typical exercise runs: "Write an equation using the variables S and P to represent the following statement: 'There are six times as many students as professors at this university.' Use S for the number of students and P for the number of professors."

Clement, Lochhead, and Monk report that "On a written test with 150 calculus-level students, 37% missed this problem and two-thirds of the errors took the form of a reversal of variables such as $6S = P$. In a sample of 47 nonscience majors taking college algebra, the error rate was 57%."

It is tempting to jump to the conjecture that these failure rates result from quick and careless misinterpretations of the wording of the problem. The investigators show, however, through detailed interviews and through altering the form of the problem, that the reversal is systematic and highly persistent. For example, the reversal is observed in problems that call for translation from pictures to equations or from data tables to equations.

Two principal patterns of incorrect reasoning emerged in the interviews: (1) Some students appeared to use a word order matching strategy by simply writing down the symbols $6S = P$ in the same order in which the corresponding words appear in the text. (2) In the second approach, students were fully aware of the fact that there were more students than professors and even drew pictures showing six S 's and one P . They still believed, however, that the relationship was to be represented by $6S = P$, apparently using the expression $6S$ to indicate the larger group and P to indicate the smaller. In other words, they did not understand S as a *variable* representing the *number* of students but rather treated it as a label or unit attached to the number 6 as in 6 feet or 6 meters; that is, they were reading the equation as they would read the statement 6 m = 600 cm, a statement of *equality* which, incidentally, should be sedulously avoided for this as well as other reasons (cf. Section 3.23). (Note that the symbols m and cm do *not* stand for variables in the latter "equation.")

The very widespread occurrence of this difficulty is confirmed by Lochhead (1981) in his report of results of giving such tasks to university faculty mem-

bers and high school teachers. Again, this was not a matter of quick and careless misinterpretation. The task was administered in written form, and the subjects gave written explanations of their reasoning. The task in this instance was “Write one sentence in English that gives the same information as the following equation: $A = 7S$. A is the number of assemblers in a factory; S is the number of solderers in the factory.”

Among university faculty members, 12% of a group in the physical sciences, 55% of a group in behavioral and social sciences, and 51% in a category “other” gave incorrect interpretations, reversing the meaning of the equation. Among the high school teachers, error rates in the same categories were 28%, 67%, and 47%, respectively. Although this was not a controlled or randomized experiment, the results testify eloquently to the persistence of the difficulty and to the fact that many individuals are not helped to overcome it in the course of their schooling.

This is a disability that should not be brushed off or treated casually, nor can one expect to remediate it by a short preliminary exercise. The most effective procedure is to give exercises in which the interpretations are traversed in both directions (words to symbols and symbols to words), and such questions should then be included on tests. The exercises should be given whenever the opportunity arises in subject matter being covered in the course, not as artificial episodes divorced from the course content.

1.16 LANGUAGE

Many aspects of the development and use of language play a deep underlying role in teaching and learning in all disciplines, not just in science. This is a huge subject attended by its own huge literature, and it is impossible to do it justice in this monograph. A few basic aspects, however, are so fundamental to our teaching that they will be mentioned here in the hope that some teachers may pursue them further in more sophisticated sources.

One aspect is that of operational definition of basic concepts. Few students, even at college level, have had direct experience, making them self-conscious about examining how words acquire meaning through shared experience. They tend to think that words are defined by synonyms found in a dictionary and, when it comes to concepts such as velocity and acceleration or force and mass, are completely unaware of the necessity of describing the actions and operations one executes, at least in principle, to give these terms scientific meaning. Since the words, to begin with, are metaphors, drawn from everyday speech, to which we give profoundly altered scientific meaning, only vaguely connected to the meaning in everyday speech, the students remain unaware of the alteration unless it is pointed to explicitly many times—not just once. Students must be made explicitly aware of the process of operational definition and must be made to tell the “stories” involved in generating numbers for velocity, accel-

eration, and so forth in their own words. This aspect is alluded to repeatedly in subsequent chapters.

The failure of many students to be aware when they do not fully comprehend the meaning of words and phrases in the context in which they occur underlies substantial portions of the "illiteracy" that we find currently deplored in many disciplines, not science alone.

Still another linguistic aspect, crucial to understanding scientific reasoning and explanation as opposed to recall of isolated technical terms, resides in the use of words such as "then" and "because." A perceptive description of the difficulties exhibited by many students is given by Shahn (1988). In connection with "then," he remarks:

[In] descriptions of many biological phenomena . . . "understanding" means mastery of a sequence such as "A then B then C then D" If, for example the letters represent stages of growth there is an obvious increase in complexity inherent in the process. Thus either omission or interchange of events signals a lack of understanding. Subsequent discussion with students [who gave incorrect answers on essay questions] showed that they really thought that the entire process was essentially equal to the sum of its parts, independent of order. It was as though in reading or hearing "then" the student was understanding "and." Now in a sense "then" does include "and," . . . but the sequential relationship is more restrictive hence more precise, and it is this distinction that many students apparently fail to grasp.

One might add that essentially the same problem frequently arises in connection with "if . . . then" statements of reasoning. Shahn (1988) also goes on to illustrations with "because":

Six true/false questions were devised which were all of the form "A because B," and which were all unrelated to biology, for example "Japanese cars are small because they use less gasoline." In each case the answer was false because either A and B were unrelated or the true statement should have been of the inverted form "B because A." Too many students answer some of them incorrectly indicating that there is indeed a problem. Generalizing from these two examples it seems that students often misread conjunctions so that they mean "and." Often "and" is part of the meaning of "because" but not the entire meaning

The problem here is not simply one of formal logic, and it is not eliminated by remedial exercises in formal logic. Although there indeed are similarities between formal logical operations on the one hand and scientific inference and explanation on the other, the processes are not identical. It is necessary

to confront the problem directly in subject matter context and to allow the students to make errors and profit from the experience.

Many teachers find it difficult to believe that college students, in particular, have difficulties such as those Shahn describes. All I can do is emphatically confirm Shahn's report with my own experience, which even applies to a significant percentage of students in highly selected groups. To convince oneself, one must try such questions with one's own students. The results are almost invariably chastening.

There are, of course, many other linguistic problems relevant to our teaching, but it is impossible to give an exhaustive discussion here. The examples discussed above have been selected from among other possibilities because of their crucial relationship to the literacy we hope to convey in science teaching.

1.17 WHY BOTHER WITH UNDERPINNINGS?

It is easy, and even tempting, to brush off the problems of cognitive development posed in this chapter by adopting the view that students who have not broken through to mastery of such basic and simple reasoning modes do not deserve additional effort on the part of faculty and staff and do not belong in introductory physics courses or even in college. The problem should be taken care of in the schools and should not be allowed to deflect and dilute the process of higher education.

Enlightened self-interest, however, if not a sense of broader societal responsibility, dictates a less callous view: A large fraction of engineering-physics students have these difficulties. They would develop a far better grasp of physics, and would develop and mature far more rapidly as professionals, if they received appropriate guidance and help at the earliest stages. Among the students who fail or who simply disappear from our courses (or who never enroll in the first place because of deep fear and insecurity in the face of quantitative reasoning) are many potentially promising minority students as well as most of our future elementary school teachers, not to speak of many others in whom improved scientific literacy would lead to the capacity for wiser leadership, wiser executive decision making, or just wiser citizenship.

The problem should indeed be taken care of in the schools, but it has not been, and will not be taken care of in the near future, because the teachers, except for a very small minority, have not developed the necessary knowledge and skills. It must be strongly emphasized that this is *not* the fault of the *teachers*. The plight of the future teachers was blindly ignored when they were in college, and they were not helped to develop the abstract thinking and reasoning skills they need in their own classrooms. The vast majority of working teachers are individuals of dedication and good will, but they will not develop the necessary reasoning skills spontaneously. They need help, and this help must be forthcoming from the college-university level in both preservice and in-service training.

Yet some university faculty, apparently without awareness of the damage being caused, pride themselves on attracting large student enrollments by offering science courses that avoid “math.” Avoiding “math” almost invariably means avoiding any and all arithmetical reasoning with ratios and division, not just avoiding use of algebra or calculus. Future teachers, if they take any physical science at all, seek out courses of just this variety. Other courses simply let them sink (or get through by memorizing without understanding), and the inevitable result is the continuing graduation of teachers who are in need of remediation the instant they graduate.

If we wish to remove from the college domain the reasoning problems described in this chapter, we—college and university faculty—must, for the time being, accept the necessity of helping students (and in-service teachers) develop underpinnings such as those described. Until this obligation begins to be discharged, we shall simply continue putting the same degenerative signal into what amounts to a feedback loop and will not be relieved of the problem at the college level.

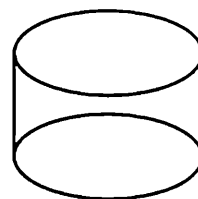
1.18 EXAMPLES OF HOMEWORK AND TEST QUESTIONS

1 Suppose we make a saline solution by dissolving 176 g of salt in 5.00 L of water. (The resulting total volume of the solution is very nearly 5.00 L.)

- Calculate the concentration of the solution, explaining your reasoning briefly.
- Using the result obtained in part (a), calculate how many cubic centimeters of solution must be taken in order to supply 10.0 g of salt. Explain your reasoning briefly.
- Make up a problem that involves the density concept and in which the steps of reasoning are exactly parallel to the steps in (a) and (b) above. Be sure to select reasonable numerical values for the physical situation you describe. Present the solution of the problem, explaining the steps briefly.

2 We have a cylindrical container A as illustrated in the figure. A second container B has the same shape as A, but the length scale, in all three dimensions, is larger by a factor of 1.80.

Answer the following questions by using appropriate scaling ratios only. There should be no appeal to formulas for areas or volumes of special shapes. Evaluate final results in decimal form. Explain your reasoning briefly in each instance.

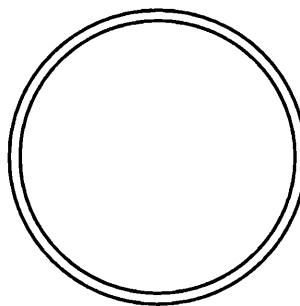


- How will the circumference C of container B compare with that of container A, that is, what is the numerical value of the ratio C_B/C_A ?
- How many times larger is the cross-sectional area (i.e., the area of the base of B, denoted by S_B) than the cross-sectional area S_A ?

- (c) If A contains 25.0 L of water when filled to the brim, how many liters of water will B hold when similarly filled?
- 3 A replica is made of the statue of a man on horseback. The total volume of the replica is 0.51 the volume of the original.
- (a) How does the length of the man's arm in the replica compare with the length of the arm in the original?
- (b) How does the total surface area of the replica compare with the total surface area of the original?

4 The earth has an equatorial radius of 3963 mi. (There are 5280 ft in one mile.) Imagine a string wrapped around the equator of a perfectly smooth earth. Suppose we now add 15 ft to the length of the string and shape the longer string into a smooth circle with its center still at the center of the earth.

How far will the string now stand away from the surface of the earth? (Be sure to make the calculation in the simplest and most economical way; avoid doing irrelevant calculations and using irrelevant data. The sketch of an appropriate straight-line graph can be more helpful than a stream of words in explaining your line of reasoning.)



5 Consider a bob on a string in uniform circular motion in a horizontal plane. Suppose that the tangential velocity v_t of the bob is increased by a factor of 2.35 while the radius of the circle is increased by a factor of 1.76. The mass of the bob remains unchanged at 145 g.

How does the final centripetal force F_{cf} acting on the bob compare with the initial centripetal force F_{ci} ?

In showing your line of reasoning, use the language of *functional variation*: for example, in geometrical situations we argued that the area factor “varies as the square of the length factor”). It is *not* appropriate to substitute the given numbers directly into a formula since the numbers are *ratios* and are not themselves velocities or radii. Avoid using any data that might be irrelevant.

6 It is an empirical fact that the power output required of the engines of a boat or ship varies roughly as the cube of the velocity, that is, if you wish to double the velocity of the boat, you must increase the power output by a factor of eight.

Consider a twin-screw boat with a mass of 2.0 metric tons (one metric ton is equivalent to 1000 kg or 2200 lb). The boat is moving at an initial velocity v_i . The captain increases the power output of the engines by a factor of 2.6.

By what factor does he increase the kinetic energy of the boat, that is, how does the final kinetic energy $K.E._f$ compare with the initial kinetic energy $K.E._i$? (Explain your reasoning briefly; use the language of functional variation, not formulas; avoid using irrelevant data; evaluate the final numerical answer in decimal form, do not leave an unevaluated expression.)

Chapter 2

Rectilinear Kinematics

2.1 INTRODUCTION

In *The Origins of Modern Science* the historian Herbert Butterfield remarks:

Of all the intellectual hurdles which the human mind has confronted and has overcome in the last fifteen hundred years the one which seems to me to have been the most amazing in character and the most stupendous in the scope of its consequences is the one relating to the problem of motion.

The Greeks with all their intellectual sophistication and mathematical skill failed to invent the concepts of velocity and acceleration, failed to grasp the notion of an instantaneous quantity and hence failed to penetrate to the law of inertia. Ideas of motion were continually belabored in the intervening years, but the breakthrough to formation and control of the concepts did not take place until the seventeenth century. This is a measure of the subtlety of the concepts and the justification for Butterfield's dramatic assertion; yet, we expect our students to assimilate the whole sequence from two or three pages of cryptic text and a rapid lecture presentation. It should not be surprising that research indicates that very few students do master the basic kinematical ideas in the first years of introductory physics [Trowbridge and McDermott (1980,1981)].

This chapter explores some of the reasons behind the existing failure and offers a few instructional strategies that might help students.

2.2 MISLEADING EQUATIONS AND TERMINOLOGY

A very common introduction to kinematics runs as follows: Suppose an object travels a distance d in an arbitrary time interval t . We define the average speed (or velocity) v by

$$v = \frac{d}{t} \tag{2.2.1}$$

Subsequently acceleration a is introduced in a similar way as

$$a = \frac{v}{t} \quad (2.2.2)$$

and eventually equations such as

$$d = \frac{1}{2}at^2 \quad (2.2.3)$$

or

$$d = v_0t + \frac{1}{2}at^2 \quad (2.2.4)$$

make their appearance. Equation 2.2.4 is then casually applied to cases of positive initial velocity and negative acceleration (e.g., throwing a ball vertically upward) in which the motion is not monotonic.

Such presentations are very misleading and essentially incorrect in certain very fundamental ways. In Eqs. 2.2.3 and 2.2.4, the symbol t does not denote an arbitrary time *interval* as it does in Eq. 2.2.1; it denotes clock readings (instants) measured from a zero setting. The symbol d in Eqs. 2.2.3 and 2.2.4 no longer denotes a distance traveled by the body; it denotes position numbers located as distances from some arbitrary origin, a point through which the body may never have passed. The students are not informed that the meaning of the symbols was changed in the derivations that followed Eq. 2.2.1, and many emerge with little understanding of either the physical concepts of velocity and acceleration or of the algebraic equations. They are hence forced to take refuge in memorizing calculational procedures that lead to “correct” numerical answers in end-of-chapter problems .

The presentation outlined above *must* be altered if students are to start kinematics with some hope of understanding the scheme. The shortcuts, omissions, and “simplifications,” which are meant to reduce “complexity,” are *not* in fact conducive to understanding; they are specious, and they make genuine understanding extremely difficult.

The concept of acceleration is inextricably connected to instantaneous velocity. It is impossible to deal clearly and correctly with instantaneous quantities without discriminating between instants (or “clock readings”) and time intervals. It is impossible to deal with back-and-forth motion without discriminating between positions, changes in position, and distances traveled by the body (three *different* concepts to which the term “distance” is frequently indiscriminately applied). These are indeed sophisticated ideas; that is why it took the human mind so long to penetrate them. It is unrealistic to expect students to make the penetration in the short time and through the shortcuts that are so frequently imposed.

2.3 EVENTS: POSITIONS AND CLOCK READINGS

The simplest and most realistic way to lead students into the kinematical concepts is to start with the concepts of “position” and “clock reading.” (This, incidentally, paves the way from the very beginning for the notion of “event,” which is so useful in introductory relativity.)

One can, for example, start with a rolling ball or moving cart in the laboratory table; make (or imagine) a “flash picture” that shows the object at uniform time intervals; place a scale behind the object; lead the students to see that the numbers on the scale do *not* represent distances traveled by the *object*; that, as distances, they are distances from an arbitrary origin at which the object may never have been located; that it takes two such numbers to give information about a change of position within a specified time interval; that we give such numbers the name “position numbers.” (In my own classes, I usually have the students sketch hypothetical strobe pictures of their own as I lead them Socratically through the above sequence.)

Students must be led to see that a number on the position scale gives the location of a reference *point* on the moving object—the distance of the reference point from an arbitrary origin at which the object may never have been located—and that a position, being defined as a geometrical point, has zero length in its own right, just as an instant of time has zero duration.

We can now introduce a clock into our picture and associate each position of the moving object with a simultaneous clock reading. A clock reading is *not* a time *interval* any more than a position is a distance travelled by the body. A clock reading is analogous to an object position (it literally *is* the position of the hand of the clock if, for the precious moment, we avoid the digital world); it takes two clock readings to make a time interval; one of the two readings may have been the zero reading but not necessarily. A given object position and the corresponding clock reading are inextricably connected, and we call the combination an “event.”

Now it becomes appropriate to couple the concept of “clock reading” with that of “instant.” This must be done carefully and explicitly because the word “instant” is being taken out of everyday speech and given an unfamiliar meaning. To most students the word “instant” means, very reasonably, a short time *interval* as, for example, “I shall be there in an instant.” They should be led to understand that, just as positions have zero length, by definition, so clock readings or instants have zero duration by definition.

If we use, say, the symbol s for position and (unfortunately but conventionally) the symbol t for clock reading, we should avoid referring to s values as “distances” or to t values as “times.” To the student the latter term invariably implies time *interval*. It is wiser and more effective to encourage use of terms such as “position” and “clock reading” (or “instant”); otherwise linguistic clarity is significantly compromised.

2.4 INSTANTANEOUS POSITION

If one carefully introduces the concepts of position and clock reading as outlined above, it is immediately possible to capitalize on this treatment by giving it deeper meaning and anticipating the more difficult notion of instantaneous velocity. If an object is moving continuously, how long does it stay at any one position number? This is not a trivial question, and most students have considerable difficulty with it. One must help them develop the following ideas.

The reference point we are using on our object is located at a particular position number *at* a corresponding clock reading (not *for* a clock reading; to the students the word “for” immediately implies a *finite* duration); how many seconds does the reference point spend at this position? (Many students will answer to the effect that the object spends a very short time *interval* at the given position.) How many seconds does a clock reading last? (Many students will again reveal their belief that the term represents a very short time *interval*, a very small number of seconds, despite having been through the development outlined in the preceding discourse.) We use the word “instant” as synonymous with “clock reading.” How long does an instant last? And so forth. Students must be led to comprehend a clock reading, or instant, as lasting for *zero* seconds and the position as being occupied for zero seconds. It is important that they *say* these things themselves; for many students it is not enough to hear them said by someone else.

This sequence brings students their first exposure to the notion of an instantaneous quantity: instantaneous position. The notion is subtle and not easy to absorb, but it is considerably easier to absorb than “instantaneous velocity.” Paving the way by introducing “instantaneous position” first makes the subsequent introduction of “instantaneous velocity” a recycling of the concepts “instant” and “instantaneous,” and this significantly reduces some of the subsequent difficulties with “instantaneous velocity.”

2.5 INTRODUCING THE CONCEPT OF “AVERAGE VELOCITY”

The most common way of introducing “average velocity” is by a statement to the effect that “average velocity over a given time interval is the change of position divided by the time interval over which the change occurred.”

$$\bar{v} \equiv \frac{\Delta s}{\Delta t} \tag{2.5.1}$$

There is nothing logically wrong with this, but starting the development of the concept with the phrase “average velocity *is* . . .” leaves most students with the impression that the name “velocity” comes first as some kind of

primitive they should "know" ahead of time, and that the idea embodied in $\Delta s/\Delta t$ comes afterwards. Teaching is significantly strengthened, however, if one carefully abides by the precept "idea *first* and name *afterwards*," not just in this instance, but in the introduction of *every* new concept. The following approach is more effective than starting with the name:

Having first generated the position versus clock-reading description of the behavior of a moving object, an effective next step is to raise the question as to how one might now devise a calculation with s and t numbers the result of which carries direct information concerning how fast the object was moving. This helps motivate examination of the ratio $\Delta s/\Delta t$, without invoking a name, but interpreting its significance by using specific numerical examples of motion of the given body along a position scale: the number is large when the object moves rapidly; the number is small when the object moves slowly; the algebraic sign indicates the direction of motion, and so on. After the utility and meaning of the number are firmly established, it is convenient to give it a name, and the conventional name is "average velocity." Then one can stand back, explicitly indicate that the concept has been introduced in accordance with the precept "idea first and name afterwards," and explain why the precept is invoked.

A very effective contrast can then be provided by asking students to examine the ratio $\Delta t/\Delta s$. How does this number behave? Under what circumstances is it large? Under what circumstances is it small? What might be an appropriate descriptive name for this quantity? Allowing the students to invent a name impresses on them the fact that the initial idea is more significant than the name and that the idea comes first. (Geophysicists give this quantity the name "slowness"; it is useful in that science because the reciprocal of velocity arises automatically in connection with the use of Snell's law in ray tracing of acoustic and seismic wave propagation.)

This approach immediately confronts students with the fact that scientific concepts are not objects "discovered" by an explorer but are abstractions deliberately created or invented by acts of human intelligence. (The same point is to be emphasized later in connection with the invention of the concept of "acceleration.")

This approach also allows a clear introduction to the notion of operational definition. Students should be led to articulate the entire "story" of the operations that go with the invention of "average velocity": creating the ideas of position and clock reading, observing two events with their corresponding values of s and t , calculating change of position Δs and the corresponding time interval Δt , dividing Δs by Δt , interpreting the physical significance of the result and giving it a name. Very few students have ever encountered the idea of careful operational definition; to most of them "defining" a term means seeking out a synonym or memorizing a single pat phrase. They are initially very resistant to the idea of telling the entire story, describing every action that goes into the creation of a physical concept. Lecture presentation, however lucid, does not make the point. The concept of operational definition is

registered only if students have the opportunity to write out such paragraphs of description in their own words and to have the writing evaluated for scientific precision and correctness of English usage.

The concept of velocity is usually introduced in connection with the simplest case, namely, uniform motion. This is proper and desirable, but textbooks and teachers frequently overlook the fact that many students do not really know what the word “uniform” means in this context. It is a familiar English word, and students pass over it without thought as to the need of translation and interpretation. They should be asked what it means and should be led to descriptions such as “equal change in position in each succeeding second.”

2.6 GRAPHS OF POSITION VERSUS CLOCK READING

Graphs of position versus clock reading are exploited to some degree in most texts. They offer a valuable alternative or supplement to verbal and algebraic treatments, offering students another way of manipulating the concepts being developed. Such graphs are most frequently (and very appropriately) used to provide a view of average velocity as the slope of a chord on the graph and to introduce instantaneous velocity as the slope of the tangent at a particular clock reading. They are also effectively used (along with velocity versus clock reading graphs) to assist the derivation of the kinematic equations for uniformly accelerated motion.

Unless they are explicitly led to do so, however, students do not consciously connect the graphs with actual or visualized motions; they treat them as uninterpreted abstractions. This is especially true of students who are still using concrete rather than formal patterns of reasoning (in the Piagetian sense of the terms). An effective way of reaching many students who have this difficulty is to lead them through direct kinesthetic experience, giving them problems in which they must translate from the graph to an actual motion and from an actual motion to its representation on a graph [See McDermott, Rosenquist, and van Zee (1987).]

The very simplest way of doing this is to give the students a set of s versus t graphs (and, eventually, v versus t graphs), as illustrated in Questions 1 and 2 among the samples at the end of this chapter, and ask them to execute the indicated motions with their own hand along the edge of the table. The reverse line of reasoning involves observing an actual rectilinear motion and sketching corresponding s versus t and v versus t graphs.

It is now a commonplace that such kinesthetic experience is conducive to firmer and more rapid mastery of the kinematical concepts, but this was not always the case. When I pointed out the significance of kinesthetic experience at a meeting in 1965 (having illustrated the use of such problems in my first textbook [Arons (1965)]), I was cut off by the meeting chairman, who said that this was merely a “personal gimmick” that was not important in imparting

physics to the students. The put-down was so forceful, and lack of interest in the audience so palpable, that I decided not to risk more of the same by trying to publish the idea apart from my text. I did, however, make use of the same questions in my second text [Arons (1977)].

Since that time, researches in teaching and learning have made clear the efficacy of kinesthetic experience, and the mode of instruction is widely accepted. Not only does the literature abound with discussions of the basic idea of making the translations in various ways, but there exist laboratory equipment and computer-based software that provide powerful assistance. The sonic range finder, in particular, has proved to be an invaluable tool. [See, for example, Thornton (1987a) and (1987b), Thornton and Sokoloff (1990), Pfister and Laws (1995), Trowbridge ("Graphs & Tracks")].

The direct, kinesthetic sensations attendant upon such exercises, as well as the thinking involved in making the translations to and from graphs, help register the concepts through use and experience. Such exercises are best done qualitatively, forming graphs and describing motions without use of numerical values.

Sophisticated equipment and computers are very fine when available and help engage the interest of the students, but I can assure readers that when the elaborate facilities are not available, the hand along the edge of the table and the translation of observed motions into graphs are still highly effective modes of instruction.

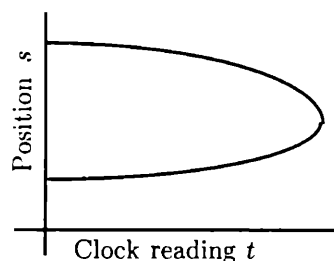
An especially important exercise with graphs is one in which students are asked to give verbal interpretations of various *lengths* in an s versus t diagram. For example, they should be able to identify a length parallel to the s -axis as representing a *change* in position. Similarly, they should be able to identify a length parallel to the t -axis as a time *interval*. A final very important contrast is provided by asking about the interpretation of a *diagonal* segment in an s versus t diagram. The majority of students do not initially have the courage of conviction to say that such a segment has no physical interpretation; they accord some spurious interpretation, most frequently a distance traversed by the body. Full understanding resides not only in knowing what something means, but in also knowing what it does *not* mean, and such exposure must be provided by the teacher (it is virtually never provided in the texts).

Another useful type of problem, rarely occurring in the texts, is that in which one examines the simultaneous behavior of two cars, say, moving at different uniform velocities and having different positions at some initial clock reading. In the light of given information, will one car pass the other? If so at what position and at what clock reading? Such problems should be solved *both* graphically and algebraically, not just in one mode; they provide a review of very basic ideas from ninth grade algebra and at the same time connect these ideas with a familiar physical situation. The great majority of students in introductory physics courses are very much in need of such review. Even many in calculus-based courses have severe difficulty setting up

the simultaneous equations.

Still another question that initially offers great difficulty is the interpretation of a graph such as that in Fig. 2.6.1. Not having had such an opportunity before, few students have the courage to say that such a representation is meaningless; they need the opportunity to say that it is meaningless and to explain *why* it is meaningless. Such experience helps them acquire security for identifying nonsense or irrelevance on other occasions. (Sample problems of the type described in this section are to be found in the last section of this chapter.)

Figure 2.6.1. Opportunity to interpret a meaningless s versus t graph.



2.7 INSTANTANEOUS VELOCITY

Acceleration cannot be carried beyond the level of being a protoconcept without engagement with the idea of instantaneous velocity. Many texts, particularly algebra-based ones, try to dodge this issue in the hope of making things “easier” for the students. The result is a specious treatment that cannot possibly lead to any genuine understanding of free fall or the law of inertia or the concept of force. Such treatments force students into memorizing calculational procedures and verbal routines that hold no meaning for them, and the result is an inevitable alienation from the subject.

I wish I knew some magic way of inculcating the concept of instantaneous velocity with no intellectual effort required from a passive student. That there is probably no such way is indicated by the long history of evolution of the motion concepts. It is by no means necessary to develop the calculus and the concept of “derivative,” but students must be given the chance to encounter the idea of instantaneous velocity slowly and with several episodes of cycling back to reencounter and reaffirm it as one proceeds through the study of kinematics and dynamics. Only a few students will absorb the concept on first encounter, but additional numbers break through in each subsequent episode. I wish to emphasize most strongly that I am not suggesting that one must stop and thrash around the concept of instantaneous velocity without moving on until every student has mastered it. This is both futile and impossible. Mastery develops slowly as the concept matures in the mind through use and application. The rate is very different with different learners. The cryptic

stand. Some slowing up is essential, together with supplementation of the kind outlined in the preceding sections.

Starting with the uniform velocity case and the corresponding straight-line s versus t graphs, one can move to cases of speeding up and slowing down with corresponding curvature of the graphs, examine chords on the graphs and their connection to average velocities over arbitrary time intervals, and finally go to the tangents to the graphs at different clock readings. The slope of the tangent can be interpreted as that uniform velocity at which the object would continue moving if change ceased abruptly at that clock reading. The slope can also be connected in the minds of students with catching the reading of the moving speedometer needle in the car at the clock reading in question. (Merely referring to the speedometer needle is not enough. Students must be led to describe what the car is doing when the needle is stationary at a nonzero reading. Then they must describe what the car is doing when the needle is swinging clockwise or counterclockwise. Then they must be led to interpret the reading caught at a particular clock reading.)

I again strongly urge that the above inquiry be conducted, and the numbers examined without introduction of the name “instantaneous velocity.” The latter term should be brought forth only after the concept has been created and the name becomes a response to the demands of convenience.

Locutions about velocity “for” an instant should be carefully avoided; “at” an instant is far more helpful and appropriate. The concept of “instantaneous position,” developed earlier, can be invoked as a helpful comparison and a review of the notion of an instantaneous quantity.

Once the concept of instantaneous velocity is established, students should be led to precise articulation of an operational definition, describing all the actions and calculations that go into obtaining the number given this name. It should also be strongly emphasized that we have created a new concept, differing from “average velocity,” even though the word “velocity” still appears. Students should be made explicitly aware of the process of redefinition that goes on continually in the creation and refinement of physical concepts. Such conscious awareness helps increase their security in the face of shifting meanings of technical terms. With “velocity,” of course, another big shift occurs when we make the transition from rectilinear to two-dimensional motion.

If students are not led to give verbal interpretation of the velocity concepts, many of them continue to regard v as an abstraction to be manipulated in formulas and replaced by numbers rather than as something intuitively comprehensible. A first stage involves getting students (particularly those having trouble) to address a question such as “What does the term ‘uniform velocity’ mean? What information does it give us about the behavior of the moving object? Tell me in simple, nontechnical words of everyday experience.” Some students will try to regurgitate the operational definition. Some will say something to the effect of “how fast it goes.” Others will flounder around with various versions of the meaningless locutions about ratios discussed in

Sections 1.8 and 1.9. Still others will talk about distance, or even position change, “over” time. (The latter locution is likely to be a trap for the teacher. The majority of students who use the word “over” are not thinking of the ratio, as one might like to believe. They are using it in the sense of “during” without conscious connection to the unit time interval.) One must persist until the student indicates that the number tells us how far the body goes in one second (or whatever time unit happens to be relevant). The “one” must be given firm emphasis; if it is hesitant or concealed, understanding is lacking.

Following this initial sequence, “average velocity” can be interpreted as that *uniform* velocity at which the object would have undergone the the same position change in the same time interval, and “instantaneous velocity” can be interpreted in the manner outlined earlier in this section. Each time it must be reemphasized that the number refers to what happens in one second.

Finally, the student must be led to see the distinction between the operational definition and the interpretation and must be helped to recognize that the interpretation, although helpful to our thought processes, does not constitute an adequate definition.

2.8 ALGEBRAIC SIGNS

If the course is one in which the full algebraic treatment of rectilinear motion is to be developed, it becomes important to lead students to see how the algebraic signs arise in the first place. However obvious it might be to us that the signs come from our uniting the number line with the position scale, this is not an insight that the students perceive or articulate spontaneously. They should be led to articulate in their own words that the algebraic sign that appears with velocity is determined by Δs while Δt is intrinsically positive. They should then articulate the fact that the sign of Δs arises because of our introduction of the number line, and that we are therefore responsible for interpreting its meaning.

It is this personal responsibility for interpretation that most students do not discern. Without examination of the origin of the signs, they memorize the interpretation as an edict from text or teacher. This lack of insight subsequently almost completely blocks interpretation and comprehension of the algebraic signs of Δv , and the blind memorization continues.

2.9 ACCELERATION

There are still some authors who seem to think that life is made “easier” for the student by introducing acceleration as $a = v/t$, apparently failing to realize the confusion caused by using the same symbol v for either an instantaneous velocity or a change in instantaneous velocity. Fortunately, this treatment is now relatively rare, and most texts recognize the necessity of dealing with

a change from one instantaneous velocity to another between corresponding clock readings. Thus, one now normally deals with some version of $\bar{a} \equiv \Delta v / \Delta t$.

As with average and instantaneous velocity, I again urge adherence to the precept “idea first and name afterwards.” Inquiry can first be directed toward devising a way of describing how fast velocity changes. The properties and behavior of $\Delta v / \Delta t$ can be examined *first* and the name “acceleration” introduced *after* the meaning and usefulness of this ratio become apparent.

It takes many students, including ones in engineering-physics courses, a long time to begin to absorb some of the physical meaning. Understanding of the acceleration concept is *not* assured by the production of “correct answers” in the conventional end-of-chapter problems, and students having trouble with such problems are almost invariably unable to describe the meaning of “acceleration.”

If asked to describe, in simple, everyday words, what “acceleration” means, many students respond “how fast it goes,” with no very clear antecedent for the pronoun “it.” If then asked to describe what “velocity” means, they give the same response. Some are surprised and a little troubled by the redundancy; others seem not to notice it. An effective approach is to go back to experience in an accelerating car and ask the student to invent a possible example with numerical values: select a velocity at a first clock reading; cite a possible velocity at a second clock reading. Do any of the numbers tabulated so far represent an acceleration? What must be done to obtain acceleration? Under what circumstances would the acceleration come out zero? How would you describe the meaning of the number in nontechnical, everyday language, that is, what does the number tell us about what is happening to the car? It usually takes substantial effort to lead students (especially those having trouble) to the point at which they say that the number tells us how much the velocity changes in one second.

One must be careful *not* to accept locutions such as “velocity per time” or “change of velocity over time.” The majority of students using the word “over” are not thinking of the ratio but are using the word in the sense of “during,” without explicit awareness of the connection to the unit interval. Some students interpret the statement “acceleration is the time rate of change of velocity” as “acceleration is the amount of time required to change the velocity.” They fail to think about problems correctly until they can *say* things correctly.

Again, as in the case of “velocity,” it is necessary to help students see the distinction between the operational definition and an interpretation.

Reversal of the preceding line of thought is also helpful, and even necessary, for many students: Suppose the acceleration of the car is 2.5 mi/(hr)(s) and the velocity at this instant is 20 mi/hr. What will be the velocity at the end of the next second? At the end of the following second? And so on. Many students initially fail to make the simple translation of the numerical values without turning to a formula.

without turning to a formula.

If the student has had some exposure to the phenomenon of free fall, it is useful to invoke the following: Have you worked with the number 10 m/(s)(s) in connection with free fall? (Student answers: Yes.) What does this number mean, that is, what does it refer to or describe? (Student frequently answers: Gravity.) The word “gravity” refers to the whole phenomenon of attractive interaction between material objects. This number cannot possibly be “gravity”; what kind of *quantity* is it? Does it have any relation to kinematic concepts we have defined? [In this way, the student may finally be led to recognize 10 m/(s)(s) as an acceleration.] Now suppose we drop a ball from rest from a high position. What will be its velocity at the end of one second? At the end of the next second? The next? And so on. Suppose we throw a ball vertically upward and it leaves our hand with an instantaneous velocity of 30 m/s . What will be the velocity at the end of the first second? The next? The third? The fourth? The fifth?

Through sequences such as this, students make steps toward a grasp of meaning of the concept, steps not induced by the end-of-chapter problems. It should be clearly understood that I do not decry, or wish to eliminate, the problems. They are essential in the learning sequence, but they are not sufficient in themselves. They must be supplemented by the induction of phenomenological thinking of the variety being illustrated.

Again, if the course is one developing the full algebraic treatment of kinematics, it is essential to pause and help the students unravel the full meaning of the algebraic signs attending Δv . Unless this is done, few students ever come to understand the origin of the algebraic sign that goes with acceleration. They must be made to realize that the interpretation goes back to our introduction of the number line and is not an a priori dictum from above; that we must make the interpretation ourselves since we originated the scheme. This is best done by having them make up reasonable numbers for initial and final velocities of an object speeding up from an initial positive velocity, then slowing down from an initial positive velocity, then speeding up and slowing down from an initial negative velocity. The resulting Δv values should be listed to help reveal the pattern, and the algebraic signs should be explicitly interpreted.

I wish there were shortcuts for this exposure, but I do not know of any. The ideas are subtle and far from trivial. If ignored in the hope that penetration will occur spontaneously with passage of time, the chickens simply come home to roost later in dynamics. Most teachers are aware of the great difficulty students have with algebraic problems in dynamics: they ignore the signs; they avoid them; they treat them carelessly and incorrectly hoping to iron it all out in connection with the “right answer” at the back of the book. Seeds for this syndrome are usually planted when time is gained by avoiding confrontation with the algebraic signs of Δv . Settling the issue with respect to Δv does not remove all the subsequent difficulties with algebraic signs in

Developing the concept of acceleration provides another illustration of the fact that scientific concepts are created by acts of human imagination and intelligence—an illustration even more dramatic than that referred to in Section 2.5. Galileo's alter ego in the *Two New Sciences* puts forth two possible ways of describing change in velocity. We would recognize these as $\Delta v/\Delta s$ and $\Delta v/\Delta t$ respectively. Galileo rejects the former on grounds that are not completely sound and adopts the latter, largely because he has the powerful hunch that free fall, which is what he seeks to describe, is uniformly accelerated in the $\Delta v/\Delta t$ sense but not in the other.

This episode vividly demonstrates the role of invention and shows that alternatives are sometimes possible. Furthermore, it demonstrates that the choice is sometimes dictated by criteria of elegance and simplicity, an idea that, at this stage of the game, is very startling to the students.

2.10 GRAPHS OF VELOCITY VERSUS CLOCK READING

The utility of s versus t graphs in providing opportunity to connect abstract concepts with concrete kinesthetic experience has been discussed in Section 2.6. Much the same points can be made about v versus t graphs. Students should be led to translate such graphs into motion of their hand along the edge of the table and into verbal description. They should also translate verbal descriptions into graphs. The computer-connected sonic range finder with its real time display of the associated graphs is of powerful help in this context. [See Goldberg and Anderson (1989) for a description of learning difficulties observed among students who have been through conventional course treatments of kinematics.]

Just as in the case of s versus t graphs, students should be led to interpret the physical meaning of various line segments on the v versus t graph: A segment parallel to the v -axis represents a change in velocity. A segment parallel to the t -axis represents a time interval. A diagonal segment has no physical interpretation. On this second go-around, following s versus t graphs, quite a few students will have developed the courage of conviction to articulate the latter conclusion, and they derive considerable satisfaction and reinforcement from their ability to do so.

Some students, particularly disadvantaged students and many nonscience majors with scant experience in quantitative or graphical reasoning, have great difficulty interpreting v versus t graphs; they attempt to memorize rather than think through the problems provided. Many can be helped by alteration of the context: The ordinate can be changed to represent population growth rate; the rate of filling or emptying of a container; the rate of import of oil; and the like. The process of interpreting such graphs, especially when the rates are negative, seems to help students arrive at understanding more quickly than if confined to v versus t graphs alone. This illustrates the importance of looking at an abstraction in more than one way.

There is now the added dimension of going back and forth between position and velocity graphs. This is exploited to some degree in some texts, but rarely to the extent necessary to achieve grasp and understanding. Furthermore, these graphical operations are rarely tested for, and anything not tested for is disregarded by most students—especially those who need the exercises most. [A few sample problems are given in the last section of this chapter. For investigations of student understanding of, and difficulties with, velocity graphs, see Brasell (1987) and Goldberg and Anderson (1989).] One danger of the computer-based display with the sonic range finder is that it presents the related graphs directly, and students do not think through the connections unless they are explicitly led to do so.

2.11 AREAS

Difficulties that students have with area concepts have been discussed in Chapter 1. The study of kinematics provides a valuable opportunity to improve their understanding through application and use of the idea in a rich, substantive context.

Some texts provide a few limited exercises involving the evaluation of areas under graphs, but these are usually too limited by being restricted to rectangular and triangular cases in which students can use the simple geometrical formulas. Many students begin to appreciate the full force of the process and the meaning of the relations only when they have to evaluate the area of a figure for which no formula exists and for which they must count the squares. Again, problems of this kind are ignored in the homework unless they appear on the tests.

Dealing with areas from the earliest opportunity in kinematics opens a number of very important intellectual doors:

For students who have taken, or are taking, calculus, it provides experience with the interpretation of the concept of “integral” without the obscuring emphasis on an algorithm for evaluation of an integral. Many students come out of calculus courses with good grades and with complete blindness as to the interpretation of an integral as an area in some related context. (This strongly suggests that mathematics instruction is as deficient in providing alternative ways of thinking about a concept and in providing pauses for interpretation and reflection as is much of physics instruction.)

For students who are not taking calculus, dealing with the areas becomes a way of dealing with, and comprehending, continuous change without the calculus formalism. If exploited at this juncture, it subsequently becomes a powerful tool in dealing with impulse-momentum and work-energy in an honest, rather than in a specious, way. It paves the way, for example, for eventual understanding of what the household electric power meter is registering. It also paves the way for better understanding of the invocation of spread sheet calculations and related use of the hand calculator.

2.12 TOP OF THE FLIGHT

All teachers are familiar with the tremendous difficulty students have with situations in which instantaneous velocity is zero while acceleration is not zero: the ball at the top of its flight after being thrown vertically upward; the ball rolling up an inclined plane and back down; the pendulum at the end of its swing (although this is intrinsically a two-dimensional rather than a rectilinear problem). Students cannot bring themselves to believe that the acceleration is not zero when the velocity is zero. These situations require clear discrimination among ideas of acceleration, instantaneous velocity, and change in instantaneous velocity but, at the time these situations are first encountered, the necessary concepts have not been firmly assimilated regardless of the lucidity of text and lecture presentations and regardless of the usual end-of-chapter exercises. There is also a fundamental linguistic obstacle that is inadvertently planted by texts, teachers, and the students themselves.

The latter obstacle arises from casual use of the word “stop,” or the phrase “come to rest” in referring to the condition at the top of the flight. Describing the ball as “stopping” or as “coming to rest for an instant” is taken very literally by the students. To them the phrases mean “standing still for a while,” and they literally think of the ball as coming to rest for a finite interval of time. Under these circumstances, the acceleration would certainly be zero.

A device that, in my experience, helps unsettle this misconception and redirects the student’s thought is the following: Suppose you observe the ball, thrown vertically upward from the ground, from a platform or helicopter that rises at a uniform vertical velocity exactly equal to the initial velocity with which the ball leaves the hand of the thrower. Suppose you also release another ball (without throwing it) from the helicopter at the same instant the other ball is thrown upward. How will the two balls behave relative to you, as you are observing them from the steadily rising helicopter?

When I first tried this sequence of questioning, I expected that many students, particularly slower thinkers, would have great difficulty with the change in frame of reference, and I was not very sanguine about its promise as an instructional device. To my surprise, I found that the majority of students respond correctly and perceptively when the questions are carefully and clearly phrased. They state that the two balls would appear to behave identically for each observer. They recognize that, from the point of view of the helicopter, both balls are falling (and accelerating) all the time. They recognize that nothing special is happening—no alteration of behavior—at the instant that the ground observer perceives them to be at the top of their flight. They begin to concede that the balls do not “stop” and that acceleration is taking place all the time, even at the instant the velocity is zero from the point of view of the observer on the ground.

All this reinforces the importance of talking about velocity *at* an instant rather than “for” an instant and continually emphasizing that any given in-

stantaneous velocity lasts for zero seconds. When the student begins to say, however tentatively and uncomfortably, that the velocity at the top of the flight is zero at an instant while acceleration at the same instant is not zero, he or she is approaching a major conceptual breakthrough—a step toward deeper grasp of the nature of instantaneous quantities and a step toward firmer distinction between velocity and acceleration.

The grasp can be strengthened by repeating the numerical exercise suggested in Section 2.9 with its rich connection to the algebraic signs: If we choose positive direction upward and the ball leaves our hand with a velocity of $+30$ m/s, what is the velocity at the end of the first second? [be sure to give the algebraic sign explicitly whenever you give a number] (Student: $+20$ m/s); at the end of the next second? (Student: $+10$ m/s); the next? (Student, tentatively: 0 m/s??). For how long does the ball have this velocity? (Student: 0 seconds??). What is the velocity at the end of the next second? (Student is likely to flounder, give a number without algebraic sign and, if corrected, finally come forth uncertainly with -10 m/s???). What has been the acceleration all the time, throughout the entire history? (When the student finally comes forth with -10 m/(s)(s), a great many things begin to fall in place simultaneously.)

Finally, Mr. Brian Popp of our Physics Education Research Group suggested, some years ago, a simple, compelling experiment: While driving a car up a gentle slope, put the car into neutral and coast. At the instant of zero velocity, abruptly put on the brake. The result is a heavy jolt associated with the “jerk” (abrupt change in acceleration), and the experiment should not be performed on too steep a slope. When the same experiment is performed in coasting to the instant of zero velocity on a level road, there is no jolt at all because there is no abrupt change in acceleration. This, of course, constitutes a preview of dynamics and the concept of force and can be exploited accordingly. One can also cycle back to this experiment when studying Newton’s second law.

2.13 SOLVING KINEMATICS PROBLEMS

The usual numerical end-of-chapter problems on kinematics constitute valuable exercises for the students, and the concentration on less familiar aspects in this book implies no derogation of the problems. Quite a few texts now present the student with sensible, systematic schemes for approaching the solutions: draw a diagram of the physical situation; set up the position line, identifying positive and negative directions; translate the verbal statement into symbols so as to (1) tabulate the known quantities together with their symbols, and (2) list the symbols of the unknown quantities; select the kinematic equation that gives the most efficient solution; make the necessary calculations; interpret the results.

When the text does not provide such help, the teacher should most certainly do so, together with posted or distributed solutions exemplifying the systematic approach.

What the teacher must be fully conscious of is the tremendous resistance many students bring to utilizing the systematic scheme despite its patent power and simplicity. In my experience, the great majority of students begin to take this process seriously only if its use is required on tests and only if substantial deductions are made when it is not used. The same resistance tends to manifest itself even more strongly later on in dynamics, and it can be reduced in marked degree if firm insistence on systematic procedure begins in kinematics.

There is another, less obvious and less frequently articulated, effect of firmly requiring use of the systematic problem-solving procedure. Most students at this early stage in their development refuse to put pencil to paper, or to analyze the verbal-to-symbol transitions that are essential, until they “see” the solution as a whole. Requiring that they institute the procedure propels them, willy nilly, into the problem, and the momentum thus acquired frequently carries them through to the solution. The increasing satisfaction gained from such experiences gradually makes them more willing to penetrate a new problem, with pencil and paper and inquiry, without waiting until the entire solution has been perceived. This is a very large step indeed in intellectual development and capacity for abstract logical reasoning. [For an excellent discussion of problem solving in more general terms, see Reif (1995).]

2.14 USE OF COMPUTERS

Kinematics is, of course, a rich field for early experience with numerical calculation and the development of familiarity with elementary computer programming. The field is widely exploited accordingly, and published materials are available. [See, for example, Eisberg (1976).] Use of the computer in this context, however, has instructional feedback effects that are not always explicitly recognized. When a student has to program a numerical calculation, he or she is exposed in the most intimate possible way to the arithmetic in which an instantaneous acceleration, sustained for a short time interval, produces a small change in velocity; the new velocity, sustained for a short time interval, yields a new position; the new position gives a new acceleration, and so on. (The exercise is valuable even in the case of uniform acceleration.)

Very few students perceive or absorb this sequence of arithmetical connection among the kinematic concepts when they are exposed only to the closed algebraic equations for the case of uniform acceleration or for some of the special cases of varying acceleration. Programming (or even doing a few numerical calculations by hand) proves to be very revealing and helps register the full meaning of the concepts.

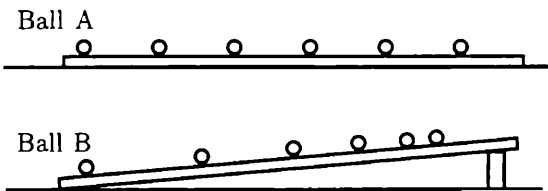
Although time is not available for every desirable activity in every course, anything that can be done to entice students into using their programmable hand calculators or home computers in this way pays dividends in improved understanding of the concepts of velocity and acceleration. There is more here than just enhancement of “computer literacy,” although I have no intention of deprecating the latter.

2.15 RESEARCH ON FORMATION AND MASTERY OF THE CONCEPT OF VELOCITY

To most of us physics teachers the concept of “velocity” (or, at least, “speed”) appears so simple and self-evident, so clearly connected with all our everyday experiences of motion, that it becomes hard to believe that students do not absorb its essentials from the usual textbook and lecture presentations. That thorough and effective intuitive grasp does not in fact develop so easily is clearly shown by the investigations conducted by Trowbridge and McDermott (1980).

In exploratory interviews, Trowbridge and McDermott found that students with no previous study of physics think of the word “speed” as a relation between distance traveled and the elapsed time but not necessarily as a ratio. Similarly, the word “acceleration” is used in a primitive sense of “speeding up” but not as a ratio. Trowbridge and McDermott describe the students at this stage as having “protoconcepts,” rather than well formed concepts, connected with the standard technical terms. They then go on to show that the protoconcept stage persists to at least some degree in many students even after formal course development of the physical concepts.

Figure 2.15.1 Speed comparison task: Motion of balls is from left to right. Ball A moves at uniform speed. Ball B starts off faster than A and slows down. There are two passing points. (See Fig. 2.15.2 for representative graphs.)

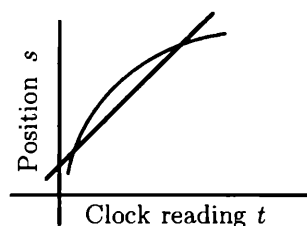


Striking illustrations of what is transpiring in learners’ minds are provided by student response to the following physical situation: The student being interviewed watches two balls rolling on parallel tracks (Fig. 2.15.1). Ball A travels with uniform motion from left to right while ball B travels in the same direction with an initial velocity greater than that of ball A. As ball B travels up a gentle incline, it slows down and eventually comes to rest. Ball B first passes ball A but, a bit later, ball A passes ball B. The student observes the motions of the balls, first separately and then together, several times and

has ample opportunity to absorb the whole picture visually. (The position versus clock reading graph shown in Fig. 2.15.2 illustrates the motions just described, but this graph was not used in the interviews.)

During the course of the interview, students were asked: “Do these two balls ever have the same speed?” (The term “velocity” was used if the student had already been introduced to it.) Trowbridge and McDermott found that a substantial number of students (up to 30% in calculus-physics courses and larger percentages in less sophisticated courses) responded to this question by identifying the instants of passing rather than the instant near which the balls maintained an almost constant separation. The association of “same speed” with “passing” or “same position” was persistent and symptomatic and not idiosyncratic.

Figure 2.15.2 Position versus clock-reading graphs for motions described in Fig. 2.15.1. (These graphs were not used in interviews with students.)



When these students watched varying motions of two balls so arranged that they did not pass each other, they said the balls never had the same speed even though there was an instant at which the speeds were indeed the same. Many students view their own experience in cars passing each other in terms of having slower speed when one is behind, faster speed when ahead, and the same speed when “neck and neck” for a “while.” (The reader interested in greater detail concerning the tasks and in direct quotations of student response should refer to the original paper.)

Trowbridge and McDermott summarize their investigation as follows:

In both pre- and postcourse interviews, failure on the speed comparison tasks was almost invariably due to improper use of a position criterion to determine relative velocity. Although students who were unsuccessful could generally give an acceptable definition for velocity, they did not understand the concept well enough to be able to determine a procedure they could use in a real life situation for deciding if and when two objects have the same speed. Instead they fell back on the perceptually obvious phenomenon of passing. Some identified being ahead or being behind as being faster or slower. We refer to this use of position to determine relative velocity as the position-speed confusion. The use of the word “confusion” here should not be misconstrued to mean the mistaking of one fully developed concept for another. We are using the expres-

sion “confusion between speed (or velocity) and position” to refer to the indiscriminate use of nondifferentiated protoconcepts. . .

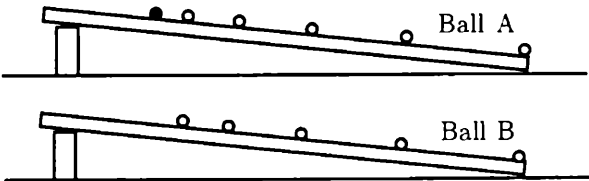
Our research also has provided evidence that for some students certain preconceptions may be remarkably persistent. . . Even on postcourse interviews, when difficulties occurred, they could be traced to the same confusion between speed and position that had been demonstrated during precourse interviews. The belief that a position criterion may be used to compare relative velocities seemed to remain intact in some students even after several weeks of instruction.

2.16 RESEARCH ON FORMATION AND MASTERY OF THE CONCEPT OF ACCELERATION

In addition to the investigation concerning the velocity concept, Trowbridge and McDermott (1981) also conducted a similar investigation with respect to acceleration.

In an exploratory sequence, students who had had some prior instruction in kinematics again viewed the motions described in Figs. 2.15.1 and 2.15.2 in the preceding section. When asked whether the two balls ever had the same acceleration, some students said the accelerations were the same when the velocities were the same. When asked how they justified this conclusion, a typical response was “because your acceleration is that Δv over Δt . And at the point where you have the same velocity, you have the same Δt and the same Δv .” These students were not discriminating between velocity and change of velocity. Further probing showed that the word “over” was being used in the sense of “during” and did not imply a ratio.

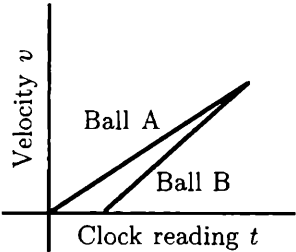
Figure 2.16.1 Acceleration comparison task: Motion of balls is from left to right. Balls roll in channels of slightly different width so the accelerations are not the same. Successive positions are shown as they would appear in a strobe photo. Shaded circle indicates initial position of Ball A. Open circles indicate corresponding positions at equal time intervals.



In a more sophisticated task, students viewed two balls rolling down inclined tracks with different accelerations. The motions they saw are described

by Figs. 2.16.1 and 2.16.2. (The different accelerations are achieved by using as tracks two aluminum channels of slightly different width, making the accelerations different even though the slopes are the same.) Ball A is released first from a point several centimeters behind ball B. After rolling a few centimeters, ball A strikes the lever of a microswitch, which in turn releases ball B. As can be seen from the graph (which was not used in the interview), the balls have the same average velocity and the same final velocity. However, ball B, which rolls on the narrower channel, reaches that final velocity in a shorter time interval than ball A and has an acceleration about 15% greater. At the base of the incline, where they achieve the same final velocity, the balls roll side by side and then enter a tunnel. (The purpose of the tunnel is to deflect attention from any subsequent, irrelevant behavior.)

Figure 2.16.2 Velocity versus clock-reading graphs of motions shown in Fig. 2.16.1. Balls reach same velocity just as they enter a tunnel at the bottom of the incline. (Tunnel was used to deflect attention from events subsequent to balls reaching bottom of incline. These graphs were not used in the student interviews.)



The balls were first rolled separately, and it was established that each one was accelerating. The students then viewed the two motions together so as to be able to compare them and were asked: “Do these two balls have the same or different accelerations?”

To encourage students to concentrate on the main conceptual issue rather than on subsidiary experimental details, specific guidance was provided. The interviewer explained that, to make the comparison, it is unnecessary to identify the cause of the acceleration or to determine whether or not the balls, the channels, or the slopes are the same. The comparison of accelerations was to be made strictly on the basis of the motions observed. It was pointed out that ball B starts later than ball A. If students did not notice that the balls entered the tunnel at the same time and did not spontaneously compare final speeds, the interviewer asked questions that served to direct attention to these aspects. Thus students were assisted in concentrating on the observations necessary for comparing the accelerations.

Trowbridge and McDermott list a hierarchy of responses that emerged, running from the most naive to those that were essentially correct as summarized in Table 2.16.1.

In pre-course interviews, only 17% of students in a calculus-physics course were successful in this task, and other groups did even more poorly—down to zero percent success in a class of academically disadvantaged students.

In post-course interviews, the success rate among the calculus-physics stu-

dents rose to 38% while that among students in two algebra-based physics courses averaged 25%. The academically disadvantaged group received specially careful instruction, not using this specific task, but addressed to encounter concrete phenomena and to improve capacity for ratio reasoning. The success rate in this group rose to 40%. (Greater detail, results with additional tasks, and information about scatter of the data will be found in the original paper.)

Table 2.16.1
Summary of Procedures Used by Students on Acceleration Comparison Task¹

<i>Procedure</i>	<i>Interpretation of Procedure</i>
1. Balls have same acceleration because slopes of tracks are the same.	Nonkinematical approach.
2. Balls have the same or different accelerations depending on their relative final positions.	Confusion between position and acceleration.
3. Balls have same acceleration because their final speeds are equal.	Confusion between velocity and acceleration.
4. Ball A has greater acceleration because it is overtaking ball B.	
5. Ball A has greater acceleration because it covers greater distance than ball B in the same time.	
6. Balls may have same acceleration because ball A covers greater distance than ball B in a longer time.	
7. Ball B has greater acceleration because its velocity changes by the same amount as the velocity of ball A but in a shorter distance.	Discrimination between velocity and changes in velocity, but neglect of corresponding time interval.
8. Ball B has greater acceleration because its velocity catches up to that of A and thus changes by a greater amount.	
9. Ball B has greater acceleration because its velocity changes by a greater amount than velocity of ball A in the same time.	Qualitative understanding of acceleration as the ratio $\Delta v/\Delta t$.
10. Ball B has greater acceleration because its velocity changes by the same amount as the velocity of ball A in a shorter time.	

¹From Trowbridge and McDermott (1981).

2.17 IMPLICATIONS OF THE RESEARCH RESULTS

These investigations dramatically illustrate the large gap that exists between the “protoconcepts” with which most students come to the study of kinematics and their grasp of the physical constructs put forth in text and lecture presentations. The investigations also show the high persistence of the gap in the face of conventional instruction.

Deficiencies in assimilation and understanding of the concepts remain concealed from us physics teachers partly because of our own wishful thinking regarding the lucidity of our presentations and partly because conventional homework problems and test questions do not reveal the true state of student thinking and comprehension. It is tempting to believe that adequate performance on conventional end-of-chapter problems indicates understanding but, in fact, it does not.

Presentations can be refined and improved to some degree, and this is always worth doing, but it is illusory to expect that vividness and lucidity of exposition are sufficient in themselves. To help the learner assimilate abstract concepts, it is essential to engage the learner’s mind in active use of the concepts in concrete situations. The concepts must be explicitly connected with immediate, visible, or kinesthetic experience. Furthermore, the learner should be led to confront and resolve the contradictions that result from his or her own misconceptions. [See Peters (1982) for additional examples with higher level students and for additional examples of useful questions.]

The gaps in understanding cannot be fully resolved for all students on the first passage through kinematics, even with better exercises and tests. Genuine learning of abstract ideas is a slow process and requires both time and repetition. Repetition without intervening time yields meager results. The most efficient approach is to move on through the subject matter but to keep returning and reinvoking the kinematical concepts in concrete, intuitive ways at every opportunity. As the ideas are reencountered in increasingly rich contexts, they are gradually assimilated—but at different rates by different individuals.

The necessary encounters must be generated through suggested observations, homework problems, and test questions that supplement the exercises prevalent in existing texts. The tasks used by Trowbridge and McDermott in their investigations are good examples; they have high instructional value. The exercises discussed in Section 2.10 play an important role. A few additional sample questions that provide such supplementation are illustrated in Section 2.20 and in Part II of this book. Teachers who explore and verify the learning problems described in this chapter will undoubtedly invent additional (and better) supplementary questions, as well as variations on the ones suggested. In doing so, they will be contributing to a pool that needs to be greatly expanded and made available in our journals and in textbooks. An instructional sequence designed to implement the insights gained in the

researches described above is outlined by Rosenquist and McDermott (1987).

2.18 GALILEO AND THE BIRTH OF MODERN SCIENCE

The study of kinematics offers an excellent opportunity to bring out certain essential features and characteristics of scientific thought by examining the intellectual thrust of the *Discourses Concerning Two New Sciences*. What is important here is *not* priority of discovery or order of development; historical insight involves elements other than chronology. That Galileo had precursors in kinematics and theories of impetus is true but relatively insignificant in an introductory course. Fruitful insight at this juncture derives from looking at what Galileo himself emphasizes in his approach:

1 Galileo was explicitly conscious of the fact that he was defining new concepts and not “discovering” objects. He argues about the alternative definitions of acceleration discussed earlier in Section 2.9.

2 Galileo very consciously and explicitly restricted the scope of his inquiry in order to master and clarify one significant issue at a time. After some discussion (in the *Two New Sciences*) of the definition of acceleration and of instantaneous velocities of bodies in free fall, Sagredo, the impartial listener, suggests that

From these considerations it appears to me that we may obtain a proper solution of the problem discussed by philosophers, namely, what causes the acceleration in the motion of heavy bodies?

and Salviati (Galileo’s alter ego) stops this line with,

The present does not seem to be the proper time to investigate the cause of acceleration of natural motion, concerning which various opinions have been expressed by various philosophers. . . . At present it is the purpose of our Author merely to investigate and to demonstrate some of the properties of accelerated motion, whatever the cause of this acceleration might be. . . .

In other words Galileo firmly rejects an Aristotelian move to provide a complete explanation of all aspects of falling motion right from the beginning of the inquiry. Salviati’s statement has a very modern stance: One of the most clearly notable characteristics of modern scientific investigation is the art of limiting the scope of inquiry in such a way as to ensure winning of one step of understanding at a time, avoiding the distraction and confusion introduced by premature or irrelevant questions. (But this procedure is, of course, not foolproof, and, in some cases, may serve to conceal important issues and inhibit solution of a problem. Deciding when and to what extent to restrict an inquiry is still the hallmark of individual genius.)

3 In “thinking away” the resistance of air to the motion of the falling body, Galileo explicitly introduces idealization into scientific thought. He recognizes that progress can be made in understanding nature without immediately dealing with natural phenomena in all their actual detail and complexity; that refinements can be developed subsequently through successive approximation. The bulk of our study of introductory physics is confined to such simplified and idealized situations, and students should be helped to remain explicitly aware of this strategy. One can hardly put the justification in more modern terms than did Galileo himself:

As to perturbations arising from the resistance of the medium, this is . . . considerable and does not, on account of its manifold forms, submit to fixed laws and exact description. Thus if we consider only the resistance which the air offers to motions studied by us, we shall see that it disturbs them all and disturbs them in an infinite variety of ways corresponding to the infinite variety in form, weight, and velocity of the projectiles. . . Of these properties . . . infinite in number . . . it is not possible to give any exact description; hence in order to handle this matter in a scientific way, it is necessary to cut loose from these difficulties; and having discovered and demonstrated the theorems in the case of no resistance, to use them and apply them with such limitations as experience will teach.

4 Galileo’s appeal to experimental evidence is frequently presented in a distorted and simplistic way by implying that the study of rolling down the inclined track was the “first experiment” and that observations and experiments were not made prior to this. Actually, there were many keen and skillful observers from classical times on down. The Greeks, for example, appealed to the resistance to compression of an inflated pig’s bladder as direct evidence for the corporeality of air, and Aristotle’s biological studies are still admired by modern biologists. The ancients, however, did not design experiments to test hypotheses. What was new in the *Two New Sciences* was the deliberate formation of a hypothesis (that $\Delta v/\Delta t$ is uniform in “naturally accelerated,” that is, gravitationally accelerated, motion) and the design of an experiment to test the hypothesis.

5 Limited by a relatively crude method of measuring time intervals (weighing the amount of water that ran out of a large container), Galileo could make observations only over a few different inclinations of the track. To reach his most significant conclusions, he had to argue to the limiting (or asymptotic) cases. Since the acceleration proved to be uniform for all inclinations at which observations were possible, Galileo argues that one would expect this behavior to persist to the limit of an inclination of 90° , at which the object would be in free fall. He thus infers that free fall must also be uniformly accelerated.

He does not confine himself, however, to only the one limiting case; he also examines the other extreme, that of the level track or zero inclination:

. . . any velocity once imparted to a moving body will be rigidly maintained as long as the external causes of acceleration and retardation are removed, a condition which is found only on horizontal planes; for in the case of planes which slope downwards there is already present a cause of acceleration, while on planes sloping upward there is retardation; from this it follows that motion along a horizontal plane is perpetual; for if velocity be uniform, it cannot be diminished or slackened. . .

Thus, by deep insight into one of the asymptotic cases, Galileo arrives at the first correct approach to the law of inertia: rather than ask what keeps a body moving, we should ask what causes it to stop.

Very few texts design situations in which students are led to think through limiting cases in order to draw insights or conclusions, or even simply to check the validity of results obtained in solving end-of-chapter problems. The situation just analyzed is one of the earliest in which students can confront such reasoning and sense its power; it is well worth exploiting for its intellectual content.

6 Neither Plato nor Aristotle believed mathematics relevant to description and understanding of the actual physical world. For Plato, uncertain physics was too far removed from the pure, abstract truth and reality of mathematical relationship: one can conceive of a line tangent to a circle, but the finest compass and straight-edge will not construct a circle and a line with but one point in common. To Aristotle, the situation seemed inverted: to him reality lay in the forms, processes, and qualities of the physical world—aspects that could never be completely described in terms of the precise, abstract, unreal truths of mathematics. This dichotomy, deeply embedded in classical learning, was carried over into the Renaissance with the classical revival. Galileo set out to overturn these views and, in the process, he initiated the prodigiously fruitful line of mathematical physics that reached towering peaks in Newton, Laplace, Maxwell, Einstein, and Schrödinger and that plays its major role in our science today.

Galileo had previously argued that the Copernican system made earth a heavenly body. Astronomy had always been a mathematical science. Since mathematics applied to the motion of the heavenly bodies, mathematics should apply to the earth. Westfall (1971) says, “If the immutable heavens alone offer a subject proper to mathematics, the earth had been promoted into that class. . . . To the mathematical science of bodies in equilibrium [Galileo] had added a mathematical science of bodies in motion.”

In the *Two New Sciences* Galileo continually propagandizes the beauty and power of mathematics and illustrates its applicability to description of natural

phenomena. After setting up what amount to the kinematical equations for uniformly accelerated motion, he asserts that he has discovered

. . . some properties of [naturally accelerated] motion which are worth knowing and which have not hitherto been either observed or demonstrated. Some superficial observations have been made, as, for instance, that the natural motion of a heavy falling body is continuously accelerated; but to just what extent this acceleration occurs has not yet been announced; for so far as I know, no one has yet pointed out that the distances traversed during equal intervals of time by a body falling from rest, stand to one another in the same ratio as the odd numbers beginning with unity.

[To develop this result from the equation $\Delta s = (1/2)at^2$ is a problem well worth assigning in homework. It makes the students think about the ratios instead of avoiding them by eternal substitution in formulas. Such exercises, coming after elapsed time, help register the ideas about ratios discussed in Chapter 1.]

To Galileo the occurrence of integer numbers in the description of a pervasive natural phenomenon had deep philosophical implications, showing that nature was in some sense “mathematical” and that mathematics could be successfully applied in natural philosophy. Such occurrences of integer numbers are fascinating to this day, whether it be in instances of resonance, standing waves, the Balmer formula, or quantum mechanics, as well as in the chemical Law of Multiple Proportions and in Mendel’s evidence for discreteness somewhere in the genetic system.

In approaching formulation of the description of projectile motion, Galileo makes the first use of the principle of superposition:

In the preceding pages we have discussed the properties of uniform motion and of motion naturally accelerated . . . I now propose to set forth those properties that belong to a body whose motion is compounded of two other motions, namely, one uniform and one accelerated . . . This is the kind of motion seen in a moving projectile . . .

After setting up the description of projectile motion, he goes on to show that maximum range must be attained at an angle of elevation of 45° and then has Sagredo say:

The force of rigid demonstrations such as occur only in mathematics fills me with wonder and delight. From accounts given by gunners, I was already aware of the fact that, in the use of cannons and mortars, the maximum range . . . is obtained when the elevation is 45° ; but to understand why this happens far outweighs

the mere information obtained by testimony of others or even by repeated experiment.

I hope that this section effectively illustrates the tremendous richness of the context. One can present the development of kinematics not only as a significant episode in intellectual history but also as an illustration of various facets of modern scientific thought and inquiry, and one can do this at an early stage with relatively simple subject matter. The development of such insights constitutes at least one part of the “general” or “liberal” education component of a science course and, as such, it is at least as important for scientists and engineers as it is for nonscience majors. This is one component, albeit not the end-all, of scientific literacy. I contend that one of the most serious deficiencies of many introductory physics courses is the failure to incorporate an examination of such intellectual dimensions.

2.19 OBSERVATION AND INFERENCE

One aspect of abstract logical reasoning with which many students have great difficulty is that of discriminating between observation and inference. The principal reason is that they have been given virtually no practice in any of their schooling. Galileo’s experiment with rolling balls on the inclined track offers an excellent opportunity for practice in a rich, nontrivial, context.

Given an account of the whole sequence (formation of the original hypothesis, design and execution of the experiment, interpretation of the experimental results), students should be asked to analyze the sequence and identify what was observed and what was inferred. Teachers who have not asked students for such performance will be astonished by the depth and extent of confusion and by the amount of guidance and help that must be provided.

(In addition to discrimination between observation and inference, examination of the inclined track experiment affords one more valuable opportunity to deal with ratios. It should be analyzed, as Galileo analyzed it, to show that ratios of displacements from rest vary as the ratios of the squares of the corresponding time intervals and not simply by examining fit to the algebraic formula. It might seem trivial to put so much emphasis on the ratios but, unless we do this at every opportunity, we will not be helping the students overcome the grave difficulties and deficiencies described in Chapter 1.)

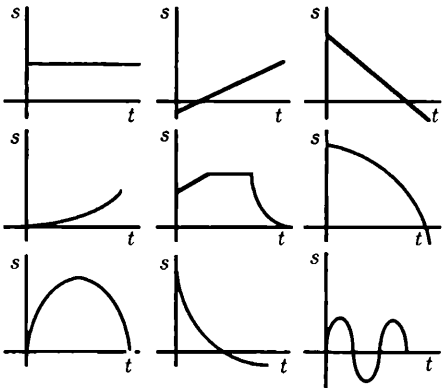
Teachers should not expect the confusion between observation and inference to be remediated in one exposure; the exercise is profitably repeated at every subsequent opportunity in other textbook contexts as well as in every laboratory experiment the students perform.

2.20 EXAMPLES OF HOMEWORK AND TEST QUESTIONS

NOTE: Problems 1 and 2 lead the student to invoke kinesthetic experience in connection with forming the concepts of velocity and acceleration and in connection with interpretation of the conventional graphs describing rectilinear motion. The use of the acoustic range finder coupled to a microcomputer in the Microcomputer Based Laboratory (MBL) materials [Thornton (1987a) and (1987b)] greatly enhances the impact of such exercises by providing immediate visual display as well as immediate feedback, correction, and reinforcement. Similar effect is to be obtained by use of Trowbridge's computer-based "Graphs and Tracks."

1 Let the edge of the table be the straight line along which motion is to take place. Think of the zero of position as being near the center of the line with positive position numbers running toward the right and negative toward the left. Let your own hand be the moving object.

Interpret each one of the position versus clock-reading histories shown in the following diagrams by performing the indicated motion with your hand. Include all the details such as speeding up, slowing down, reversing direction, standing still, moving at uniform velocity, having your hand at the appropriate position at zero clock reading and at the end of the history, and so on. Describe the motion in words as you execute it.

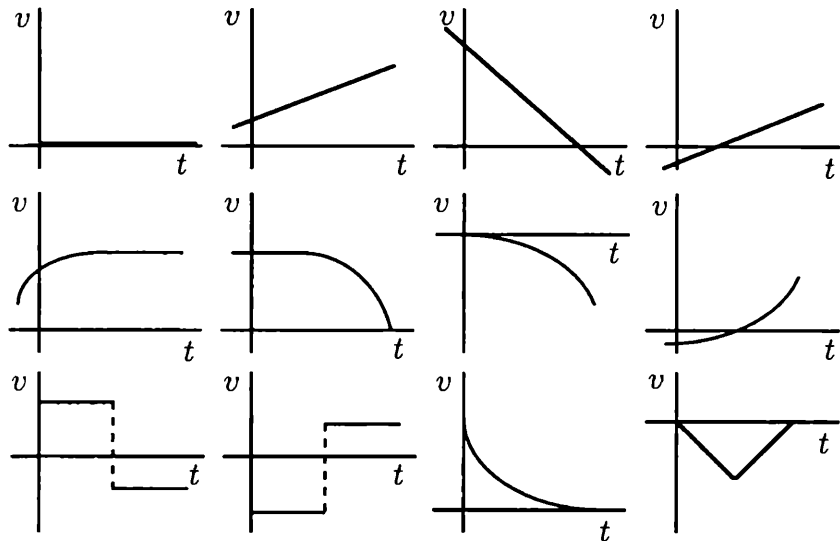


After having executed the motion with your hand, sketch the v versus t diagrams. In your sketch, be sure to place the velocity diagram directly below the position diagram so that corresponding clock readings match up.

2 Let the edge of the table be the straight line along which motion is to take place. Think of the $s = 0$ position as being near the center of the line with positive position numbers running toward the right and negative toward the left. Let your own hand be the moving object.

Interpret each one of the following velocity versus clock reading histories by executing each motion with your hand, following all details carefully as in Question 1. Does the diagram tell you where your hand should be at $t = 0$? Execute each motion more than once, each time placing your hand at a different initial position at $t = 0$. Describe the motion in words as you execute it.

After having executed the motions and described them, sketch a corresponding s versus t diagram for each v versus t diagram. Be sure to align the position diagrams directly above the velocity diagrams so that corresponding clock readings match up.



3 Cars A and B travel along the same straight road in the following manner: Car A is located at position $s = 2.4$ mi at clock reading $t = 0.00$ h and maintains a constant speed of 36.0 mi/h. Car B is located at $s = 0.0$ mi at clock reading $t = 0.50$ h and maintains a constant speed of 50.0 mi/h.

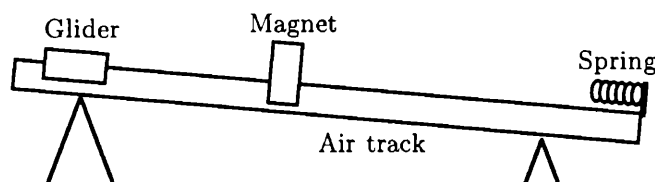
At what clock reading will car B overtake and pass car A? At what position will the passing take place? How long a time after being at $s = 0.0$ will B overtake A? At the instant of being passed, how far will A have traveled from the position occupied at $t = 0.0$?

Check yourself by solving this problem in two different ways: First solve it graphically by plotting the two s versus t histories on the same diagram and reading the required numbers off your graph. Then solve the problem algebraically by writing down two equations: one for the position s_A of car A as a function of clock reading t , and another for the position s_B of car B as a function of t . To do this, you must translate the verbal statement of the problem into symbols. You will now have two equations that you can solve simultaneously for the unknown quantities as in ninth grade algebra.

4 *Note to the instructor:* Most of the tasks used by Trowbridge and McDermott [(1980),(1981)] in their investigation of students' understanding of the concepts of velocity and acceleration can be adapted to instructional purposes, helping students master the concepts. The physical demonstrations can be set up in class or lecture and the questions asked, giving students opportunity to watch as many repetitions as they wish and to argue with each other about the answers. This is a very effective way of helping students confront the concepts intuitively, in concrete situations, and gain those insights that are not conveyed in the usual textbook problems. Trowbridge (1988) has prepared computer-based materials, under the title "Graphs and Tracks," that provide such exercises via the computer. He received a national award for these materials.

5 *Note to the instructor:* Peters (1982) describes an excellent demonstration,

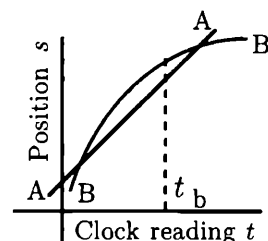
somewhat richer and more complex than those of Trowbridge and McDermott, and particularly suitable for engineering-physics courses. The apparatus is sketched in the following figure. A glider slides down a slightly inclined air track, which has a bumper spring at its lower end. A standard horseshoe magnet is placed above the middle of the track so that the glider passes between the poles of the magnet without rubbing. The glider starts from rest at the upper end of the track, speeds up, moves at uniform speed (because of eddy current effects) between the poles of the magnet, speeds up along the lower portion of the track, bounces back from the bumper spring almost up to the magnet, then returns and bounces once more.



Numerous repetitions of the motion were carried out in front of the class, and students were then asked to sketch, on a blank piece of paper, the s versus t and v versus t graphs. This exercise was given to an honors section of the calculus-physics course after rectilinear motion had been covered in class and s and v had been given precise meaning. Peters reports that only 30% of the students in the honors section represented the motion reasonably accurately on first experience with such a task. He also describes and analyzes some of the more widely prevalent incorrect responses and types of confusion that were evident.

6 Note to the instructor: The following problem is an easy-to-grade, pencil-and-paper version of the Trowbridge-McDermott speed comparison task discussed in Section 2.15. In my own experience, statistics with respect to performance on this problem are surprisingly similar to those reported by Trowbridge and McDermott for performance on the concrete task. Encounter with this problem helps some students step beyond their protoconcepts and progress toward better discrimination between position and velocity.

The figure shows position versus clock reading histories of rectilinear motion of two balls A and B rolling on parallel tracks.



- Mark with the symbol t_a along the t -axis on the diagram any instant or instants at which one ball is passing the other.
- Which ball, A or B, is moving faster at clock reading t_b ?

- (c) Mark with the symbol t_c along the t -axis on the diagram any instant or instants at which the two balls have the same velocity.
- (d) Over the period of time shown in the diagram, Ball B is (circle the correct statement among the following):
 - (1) speeding up all the time
 - (2) slowing down all the time
 - (3) speeding up part of the time and slowing down part of the time.

7 Observation to be made outside of class: Take a ball (such as a tennis ball or any child's toy) and drop it vertically from your outstretched hand. Observe the bouncing carefully several times. Then sketch s versus t , v versus t , and a versus t graphs for the observed behavior. Be sure to place the diagrams vertically below each other so that corresponding clock readings line up appropriately.

8 Observation to be made outside of class: Take a sheet of paper, hold it so that the sheet is parallel to the floor, and let it drop vertically. Observe its behavior and the character of its acceleration. Now crumple the sheet into a tight ball and let it drop vertically. Compare the two cases in your own words, describing and interpreting the differences.

9 Observation to be made outside of class (or demonstration to be performed in class): Take a string about 3 m in length and attach weights (such as metal nuts or washers, or stones, or pieces of wood) at uniform intervals of 30 or 40 cm along the string. Standing on a chair, table, or ladder, as may be necessary, hold the string of weights vertically with the lowest weight at a distance above the floor equal to the spacing of weights along the string. Let the string fall and listen carefully to the clatter of the weights as they strike the floor. Describe the sound that you hear: Does the clatter speed up, slow down, or remain uniform?

If the clatter does not remain uniform (i.e., uniform time intervals between weights striking the floor), how would you space the weights to make it uniform? Try the experiment.

10 *Note to the instructor:* Following is a type of problem that makes students confront a case of nonuniform acceleration and recognize that the available kinematic equations are not applicable. Such encounter is important and illuminating, and yet it is very rarely generated in introductory courses.

An object starts from rest at position $s = 0.0$ at clock reading $t = 0.0$. At clock reading $t = 5.0$ s it is observed to be at position $s = + 40.0$ m and to have an instantaneous velocity $v = + 11.0$ m/s.

Examine the interconnections of the given data carefully. Was the acceleration of the object uniform or nonuniform? Explain your reasoning. Are the kinematic equations you have been using in class applicable to this case? Why or why not? Sketch the shape of the velocity versus clock reading graph that is implied by the data, that is, is the graph straight or curved? If it is curved, is it concave upward or downward?

11 *Note to the instructor:* The following question requires verbal interpretation of terms in an equation. Students almost never encounter such questions, yet the

practice is an essential ingredient in learning and understanding. Similar questions should be asked in connection with the equations derived subsequently for projectile motion.

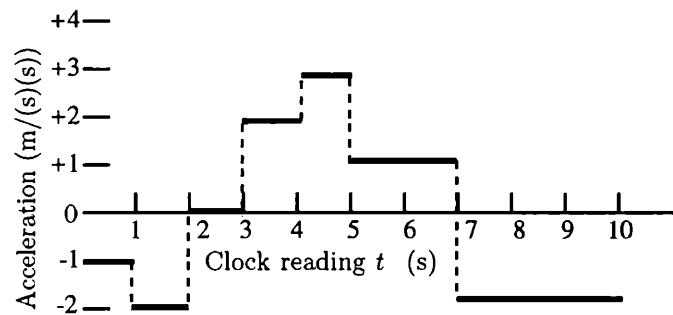
Consider the following familiar kinematical equation describing change of position with respect to clock reading in rectilinear, uniformly accelerated motion:

$$\Delta s = v_0t + (1/2)at^2$$

In your own words, give a physical interpretation of each of the two terms on the right-hand side of the equation.

12 Note to the instructor: The following type of exercise is helpful in leading students to perceive the difference between acceleration and velocity and to establish the connection between acceleration and change in velocity:

The diagram shows the acceleration versus clock reading history of a rectilinear motion. There are periods of uniform acceleration with very abrupt jumps from one acceleration value to another. This is quite possible physically. Although the acceleration changes cannot actually take place instantaneously (i.e., in zero time interval), they can take place in time intervals very short compared to the scale employed on the graph. That is what is implied in this instance.



Immediately above this diagram, on an identical time scale, plot a graph of the velocity versus clock-reading history of this motion, assuming that the body starts from rest at $t = 0$. Describe what the graph would look like if the body had some non-zero initial velocity at $t = 0$.

Chapter 3

Elementary Dynamics

3.1 INTRODUCTION

In the study of physics, the law of inertia and the concept of force have historically, been two of the most formidable stumbling blocks for students, and, as of the present time, more cognitive research has been done in this area than in any other. That the learning problem is formidable should not be surprising in view of how long it took the human mind to unravel these aspects of natural phenomena in the first place. Newcomers invariably have to relive at least some of the original hurdles and difficulties even though we shorten the time and smooth the way by providing guidance and instruction.

Most of our students come to us imbued with intuitive rules or notions that we are strongly tempted to call, pejoratively, “misconceptions.” These intuitive notions are, however, neither perverse nor idiosyncratic; they are rooted in everyday experience, and they were initially held by all our predecessors. Our pedagogical orientation becomes sounder and more reasonable if we characterize these notions as understandable “preconceptions” to be altered through concrete experience, rather than as ignorant “misconceptions” to be removed instantaneously through verbal inculcation and a few demonstrations in which the student does not actively participate.

Researches (to be cited later in the body of this chapter) have repeatedly shown these preconceptions to be very deeply rooted and highly resistant to change. Furthermore, the views held by the learner are not necessarily consistent and tend to shift from one physical situation to another, exhibiting contradictions that are not spontaneously perceived as such.

As with the kinematic concepts discussed in Chapter 2, one cannot expect the learner to acquire mastery of dynamics through verbal presentation alone, however lucid. Conventional end-of-chapter problems are also insufficient. This is not meant to disparage or advocate the elimination of such problems; they provide absolutely essential exercises in using the tools of the subject and, without them, the student would never attain the capacity to apply and use the laws of motion. In existing texts, however, most end-of-chapter

problems tend to concentrate on calculational procedures and on end results that rarely induce phenomenological, experiential thinking of the kind that research shows to be helpful in overcoming the conceptual barriers. It is shown repeatedly that ability to get correct, or partially correct, answers to the problems carries no assurance of genuine understanding of the basic concepts. End-of-chapter qualitative, phenomenological questions are also insufficient in themselves when not accompanied by concrete experience, Socratic guidance, and, eventually, testing.

Clear, vivid presentations, together with conventional quantitative problems, must be supplemented with questions and problems that engage the minds of learners in qualitative, phenomenological thinking. Learners must be confronted with direct experience, and with contradictions and inconsistencies, in such ways as to induce them to articulate lines of argument and reasoning in their own words and to lead them to abandon the deep-seated, plausible, intuitive preconceptions that impede development of the contra-intuitive but “correct” view. Most learners require several such encounters, distributed over time in increasingly rich context, and one must not expect to “rectify their disabilities” in one remedial session.

This chapter represents an effort to help the teacher become aware of some of the gaps that remain in many existing presentations and to give examples of supplemental treatments and exercises that seem to help the learner. Experience in using some of the hands-on approaches recommended is reported by Hake (1987) and by Tobias and Hake (1988) in a controlled experiment involving undergraduates as well as nonscience faculty colleagues at Indiana University.

3.2 LOGICAL STRUCTURE OF THE LAWS OF MOTION

The philosophical-epistemological basis of Newtonian Mechanics has been discussed at great length, over many years, in numerous treatises, and this is not an appropriate place to review this extensive literature. [An interested reader will find excellent summaries of modern views, relevant to physicists, in the papers by Eisenbud (1958) and Weinstock (1961) cited in the bibliography.] Before going on to description and analysis of student conceptual difficulties, however, it is appropriate to consider certain logical aspects of the laws of motion that are frequently ignored, or glossed over much too quickly, in many text presentations.

Many presentations start in by ignoring the fact that the words “force” and “mass,” which, in everyday speech, are heavily loaded metaphors, are being taken out of everyday context and given very sophisticated technical meaning, completely unfamiliar to the learner. (It is even implied, in some presentations, that the student already knows the scientific meaning of the terms.) Students have, in general, not been made self-conscious about, or sensitive to, such semantic shifts, and they continue to endow the terms with

the diffuse metaphorical meanings previously absorbed or encountered. It is helpful to make students explicitly conscious of the fact that the words remain the same but that the meanings are sharply revised.

This is a matter of operational definition, but many texts, unfortunately, either ignore operational definition entirely, proceeding as though the words have already been defined, or cryptically state a sequence that is essentially circular. The more elementary the text, the greater the tendency toward circularity and weakness of definition—apparently in the hope of making things “easier” for the learner. Given such presentations, there is no real hope of having students understand the concepts. How far one delves into operational definition of “force” and “mass”—with what degree of intensity, rigor, abstractness, and detail—is a matter of judgment for the teacher, but the matter should not go by default.

Widely different levels of sophistication are possible, and a teacher should make a choice reasonably matched to the students being addressed. Furthermore, the process of definition can be extended over time and need not be settled completely on the first encounter. One can start in some relatively unsophisticated way and help students refine the concepts by spiralling back to more rigorous definition as their grasp of the overall structure grows in later contexts.

There are two principal approaches to careful operational definition of “force” and “mass”: one I shall call “Newtonian” for lack of a better term (Newton himself never actually propounded clear operational definitions of these terms); the other is associated with the name of Ernst Mach (1893).

In Mach’s sequence, inertial mass is defined first. This is done by invoking the reaction car experiment, accepting as a law of nature the empirical observation that the ratio of the accelerations (and hence of the velocity changes) of the two bodies is a fixed property of the bodies, and defining the ratio of the masses as the inverse ratio of the accelerations. The net force acting on one body is then defined as the *ma* product for that body. [This is, of course, only a very cryptic summary of the more extensive line of argument. The reader interested in full detail will find an excellent presentation by Weinstock (1961)].

In scanning a number of widely used textbooks (I make no pretense of having carried out a full survey), I find that a significant minority use the Mach sequence. Since this sequence is basically sound and internally consistent, I shall not discuss the pedagogy in detail except to say that most of these presentations are so cryptic and so abstract that few students have any real chance of forming a sound operational grasp of the concepts from the textbook presentations. To induce such grasp, teachers would have to expand the development, give it far greater concreteness, and lead students to interpret, explain, and analyze in their own words.

Since the majority of the widely used texts adopt what I have called the “Newtonian” sequence (starting with force rather than inertial mass), and

since I am myself partial to this approach because of its greater concreteness, I shall analyze this sequence in greater detail. I hasten to emphasize, however, that I do not put forth this sequence as the one and only correct presentation. There is no one “absolutely correct” or necessary road through this epistemological terrain. What counts eventually is the internal consistency of the network one elects to form. It is up to each teacher to select the variations he or she can help students articulate most clearly and compellingly, subject, of course, to the constraints of logical consistency and absence of circularity.

[A mathematically sophisticated version of the phenomenological sequence outlined in the following sections is given by Keller (1987).]

3.3 AN OPERATIONAL INTERPRETATION OF THE FIRST LAW

The law of inertia, or Newton’s first law as most of us call it, was not new with Newton. Galileo almost had it, and Descartes did have it, right. By the time of publication of the *Principia*, the first law had become assimilated to the thinking of most active and productive natural philosophers even though, for some decades, the physics of motion continued to be taught out of scholastic texts. In the *Principia*, Newton does not arrogate the law to himself. He acknowledges the precedence of others and puts it forth as a declaration of independence from Aristotelian and impetus schools of thought.

Newcomers to dynamics, burdened with common sense ideas and rules about the behavior of moving bodies, have very great difficulty following this breakthrough, and the learning problems this entails will be discussed in later sections. Here I wish to consider only one facet of the first law: How can we interpret it operationally in the sequence of definition of concepts.

Among the list of definitions at the beginning of the *Principia*, we find the following Definition IV:

An impressed force is an action exerted upon a body in order, to change its state, either of rest, or of uniform motion in a right line.

Then, as Law I of three Laws of Motion, we find:

Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it.

The circularity here is quite apparent, but it, in fact, does suggest how we might help a student interpret Law I in our modern sequence: Up to this point, we have generated only operational definitions of the concepts of kinematics, and “force” and “mass” remain undefined. Once we begin to accept the view that rest or uniform rectilinear motion are natural states of objects

and that interactions with other objects are necessary to produce *changes* in such states, we can interpret Law I as giving us a *qualitative* operational definition of “force,” namely that action, by an agent external to the moving body, that imparts a *change* in velocity, and “change” includes both magnitude and direction.

This becomes a first step toward an operational definition of “force.” The next steps come from construction of Law II.

3.4 OPERATIONAL DEFINITION OF A NUMERICAL SCALE OF FORCE

As indicated in the preceding section, intrinsically associated with enunciation of the law of inertia we discern a *qualitative* conception of force as any action, impressed externally, changing the velocity of a body. The next step is to refine the concept by making it quantitative.¹ At this point, more than one approach is possible. Newton, in fact, elected to associate “motive force,” as he called it, with impulsive changes in momentum [for more detail on this aspect see Arons and Bork (1964)]. Our modern conception of force is different from Newton’s and it is best to carry out the discussion in modern terms.

We start by visualizing operations we could perform with frictionless pucks on a level glass table top or on an air table. (The *PSSC Physics* films on “Inertia” and “Inertial Mass,” with Edward Purcell as narrator, in fact carry out something very close to the gedanken experiments to be described. See Bibliography for current source.) Selecting a particular puck A, which becomes the standard body in our experiments, we impart rectilinear accelerations by pulling it with a light spring, the extensions of which can be observed and marked on an initially unmarked card (Fig. 3.4.1).

Intuition tells us, correctly, that different strengths of pull impart different accelerations to body A. With a particular action or pull we shall associate the numerical value of the acceleration imparted and construct what amounts to a “force meter.” Thus we imagine conducting the following experiments: (1) make a multiple-exposure photograph of accelerating puck A by flashing a

¹Here, incidentally, is an opportunity to make students explicitly aware of the fact that definitions of new concepts are rarely, if ever, generated completely, in full rigor, on the first encounter. One usually starts with an initial, tentative, even crude, definition and extends and refines it as insight deepens with use and application. This is precisely what happens with the concept of “velocity,” where we start with a notion of average speed in rectilinear motion, refine the concept by infusing algebraic directions along the number line, refine it further into the concept of “instantaneous velocity,” and finally generalize the vector properties in two and three dimensions. In each step of redefinition, the concept is altered significantly; it becomes, to all effect, a new concept even though the original name is retained. Our modes of instruction tend to lead students to concentrate on the name while losing track of the ideas behind it. It is an intellectually significant experience for the student to stand back and become explicitly conscious of the processes of definition and redefinition at such junctures.

light at successive uniform intervals of time; (2) from the sequence of increasing displacements in the photograph, we can determine whether the acceleration is uniform and whether the extension of the spring is constant.

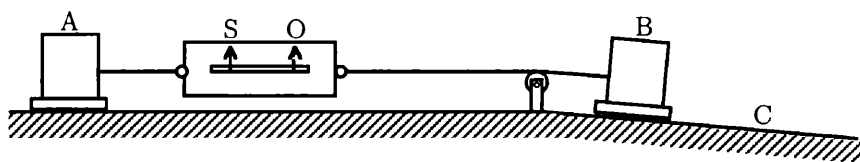


Figure 3.4.1 Frictionless puck B on incline C imparts uniform rectilinear acceleration to puck A. Acceleration can be changed by increasing or decreasing slope of incline. Needle attached to end of spring is at position O when acceleration is zero and spring is relaxed. Spring is extended, and needle is at position such as S when acceleration is imparted. (Note that *no* assumption is being made concerning spring linearity or the obeying of Hooke's law.)

All measurements of this type, whether made directly in the manner shown in Fig. 3.4.1 or accomplished in some indirect fashion, indicate that a constant spring extension is associated with a constant acceleration. Furthermore, we can satisfy ourselves that the effect is reproducible: the same spring extension imparts the same acceleration on different occasions and in different directions (right or left, north or south).² Having established confidence in the uniqueness and reproducibility of each experiment, we complete the scale of our force meter by labeling each needle position with the numerical value of acceleration imparted to puck A.

Thus, the numbers 1.00, 2.00, 3.00, and so on would be placed at needle positions under which accelerations of 1.00, 2.00, 3.00 m/(s)(s), etc., were measured on the photographs. Noninteger values would be established in a similar way: the number 1.50 would *not* be entered half way between 1.00 and 2.00 but at the needle position that imparted an acceleration of 1.50 m/(s)(s); similarly for force readings such as 2.36 or 3.82. In other words, the force scale is calibrated without *any* assumptions whatsoever concerning uniformity or nonuniformity in the stretching of the spring, that is, the spring is *not* assumed to obey Hooke's law.

If puck A is constructed to match the international standard object called

²Depending on the level of sophistication that is appropriate, one can take this opportunity to make additional, finer points: The spring must not be stretched so far that the needle fails to return to its initial, zero position at zero acceleration, but this behavior can always be checked between experiments. The care that must be exercised in calibrating the force meter is the same as that which must be exercised with clocks and meter sticks in measuring time intervals and lengths; precise measurements are to be made under conditions of controlled temperature and freedom from shock, bending, and other extraneous effects. In practice, knowledge of what effects are extraneous and how these must be controlled is rarely discerned a priori but is achieved through trial and error and successive approximations.

“one kilogram,” we give the units marked on our force meter the name “newtons.” We now have a tentative definition of force on a numerical scale. The force numbers, which we shall denote by the symbol F , have *arbitrarily* been made identical with the numerical values of acceleration imparted to the standard body, puck A. Whether this arbitrary definition of a force scale is fruitful and useful can be determined only by appeal to nature through further experiments.

3.5 APPLICATION OF THE FORCE METER TO OTHER OBJECTS: INERTIAL MASS

If we now replace puck A by a different frictionless puck, denoted by D, we can impart accelerations to D using any reading we wish on the force meter. In such experiments we find that a fixed scale reading, such as 3.00 N, on the force meter imparts a constant and reproducible acceleration to D, but this acceleration is not, in general, 3.00 m/(s)(s) as it is with puck A. Suppose the acceleration in this instance (force reading 3.00) turns out to be 1.50 m/(s)(s). Note that it is *not* possible to tell what will be observed with still other force readings; one must proceed with the experiments. With other force readings, do we obtain results systematically and simply related to the one so far observed?

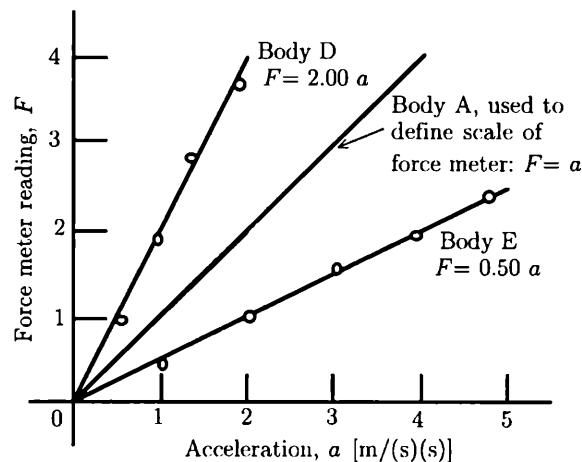
Table 3.5.1 illustrates results that would actually be obtained (column 3) and contrasts them with results that can be imagined but are not actually obtained (columns 4 and 5). Note the pedagogical importance of showing the student what is *not* the case as well as what is. Without such explicit contrast, the significance of the idea being presented is frequently unappreciated or incompletely understood. The best way for the student to grasp the idea contained in Table 3.5.1 is to sketch the F versus a graphs for the data in columns 3, 4, and 5.

Table 3.5.1
Accelerations a imparted to Body D by force readings F exerted by the force meter defined in Section 3.4.

(1) Applied force F (defined by acceleration imparted to A: units not named)	(2) Acceleration imparted to A, m/(s)(s)	(3) Observed acceleration imparted to D. m/(s)(s)	(4) Imagined possibilities of acceleration of D (not realized experimentally). m/(s)(s)	(5) Imagined possibilities of acceleration of D (not realized experimentally). m/(s)(s)
0.50	0.50	0.25	1.00	3.00
1.00	1.00	0.50	1.10	2.50
1.62	1.62	0.81	1.20	2.20
2.00	2.00	1.00	1.40	2.00
3.00	3.00	1.50	1.50	1.50
4.00	4.00	2.00	1.60	1.00

Examining column 3 in the table and the graph in Fig. 3.5.1, we see that it is possible to associate with puck D a *single* number, namely 2.00, which will, in each observation, give the force meter reading when multiplied by the acceleration imparted. Similar results are obtained with other bodies as illustrated in Fig. 3.5.1, except that the numerical factor multiplying the acceleration to give the force is different for each different body. (For body E, for example, the number is 0.50.) Thus we find, by experiment, a new law of nature: Forces are directly proportional to the accelerations imparted to bodies other than the standard one for which the force scale was arbitrarily defined, and the proportionality constant is clearly a unique value, a *property* of each new body. (Note how this treatment can be directly connected with the straight-line ideas discussed in Section 1.11.)

Figure 3.5.1 F versus a graph for bodies A, D, and E. Body D has larger inertia than A (given force meter reading imparts smaller acceleration). Body E has smaller inertia than Body A. For Body D, $F = 2.00a$; for body E, $F = 0.50a$.



Summarizing the argument: Once we have arbitrarily defined a force scale as in Section 3.4, it is found to be an experimental physical fact that F is proportional to a when different forces are applied to another body; that is, nature tells us that there exists a *single* number—a *property of the given body*—which is the proportionality constant. If we denote this proportionality constant by m , we write

$$F = ma$$

where m , the property of the body being accelerated, is the slope of the corresponding straight line in Fig. 3.5.1. We give this property the name “inertial mass” or simply “mass,” for short. The existence of this single, unique number for a given body is *not* just a matter of definition, as was the scale of force, nor is it deduced from theoretical principles; it is an experimental, physical *fact* a law of nature—even though it was originally arrived at by conjecture rather than by direct experimental test.

Having arrived at this point, one can now lead students into discussion of the meaning of large and small values of m , comparing behavior of the

bodies under action of the same force, and interpreting the significance of the fact that two entirely different bodies (different in size, shape, color, density, texture, and chemical composition) might have identical values of m , including the special value $m = 1.00$ kg.

Many students, teachers, and texts fall into the habit of using the term “mass” to denote an object, for example, speaking of “suspending a 10 kg mass.” This linguistic carelessness is the source of certain kinds of confusion, especially later on, for example, when one wishes to distinguish between gravitational and inertial mass. It is best to avoid using the terms “object” and “mass” synonymously and to distinguish carefully between an object and its properties.

3.6 SUPERPOSITION OF MASSES AND FORCES

The preceding sections have shown how noncircular operational definitions of force and inertial mass can be constructed via the second law using what I have termed the Newtonian sequence, that is, starting with force and acceleration rather than with Mach’s reaction car experiment. This, however, is still not the entire content of the second law. There remain the questions of superposition of forces and masses, and again one must appeal to experiment for verification of conjectures, however plausible the latter might be.

Experiment confirms that masses add (or subtract) arithmetically when bodies are combined (or separated). Experiment also confirms that (with the application of two identically calibrated force meters) two equal forces in the same direction impart twice the acceleration imparted by one of the forces acting alone; that equal forces in opposite directions subtract (or “cancel”)³ each other and impart zero acceleration to any object; that, in general, colinear forces superpose algebraically; that forces at angles to each other add in the same manner as velocities and accelerations, thus behaving as vector quantities; and that the *acceleration* (and *not* the velocity of the body) is always in the direction of the resultant force. (Many students confuse the latter issue, and they must be helped to make it explicit through questions on homework and tests. Such questions must usually be supplied by the teacher; they are rarely given in textbooks.)

Finally, it is an additional empirical fact that orthogonal components of velocity, acceleration, and force are independent of each other in the realm of validity of Newtonian mechanics, whereas this is not the case when relativistic effects become significant.

³One must be careful with the term “cancel” in this context. Many students tend to misunderstand and misuse it. Some have the notion that, when forces “cancel” each other, they cease to exist. Others confuse such “cancellation” with cancellation by *division* in algebra or arithmetic.

3.7 TEXTBOOK PRESENTATIONS OF THE SECOND LAW

It is unfortunate that many textbooks, in their efforts to be “simple,” or “easy,” or concise, avoid careful operational definition and completely omit discussion of what aspects of the second law involve arbitrary definition and what aspects reflect a specific kind of order in nature. Such presentations leave the students with formulas:

$$\vec{F}_{\text{net}} = m\vec{a}$$

or

$$F_{x \text{ net}} = ma_x \quad ; \quad F_{y \text{ net}} = ma_y$$

but with virtually no understanding of the content and meaning of the second law.

It is important for students to realize that the algebraic statement is not self-contained and that it must be supplemented by a fairly extended text, giving a story of arbitrary definition and appeal to experiment along lines comparable to those illustrated in the preceding sections. Without the story, the formulas are sterile and unintelligible.

Ignoring these logical and conceptual aspects of the laws of motion, in order to make things seem “easier” or to achieve more extensive coverage, shows little more than contempt for the minds of the students. Most students *can* understand these ideas if they are given time, opportunity, concrete experience, and suitable spiralling back from later context. Very few students can absorb or understand these ideas when subjected to the pace and brevity prevalent in most of our texts and courses, whether it be at high school level or in college level calculus-based or algebra-based physics courses.

In most texts adopting the Mach sequence, the presentation is made so cryptically and so abstractly as to be quite meaningless to the majority of students, even though the conceptual development is sound and not circular. The questions of superposition are rarely made explicit or given any acknowledgment whatsoever. The verbal text, the qualifications and interpretations that accompany the second law, are entirely omitted. The more “elementary” the textbook, the more cryptic and less intelligible is likely to be the presentation.

The majority of widely used textbooks seem to adopt what I have called the “Newtonian sequence,” but most of these start with “force” as though it were a primitive, already fully understood both qualitatively and numerically, and not requiring explicit operational definition. They then go on to “mass” as simply the proportionality constant between force and acceleration. The superposition questions are, for the most part, ignored.

Scanning some currently available textbooks for a few specific examples (with no pretense of complete coverage), I note that *PSSC Physics* (all editions) gives a simple, correct, and consistent presentation suitable for introductory levels. The treatment is (appropriately) less sophisticated than that outlined in the preceding sections, but it is quite reasonable for many introductory

college contexts as well as for the high school level being addressed. Among college level calculus-physics textbooks, both Tipler (1982) and Resnick and Halliday (1977, 1985) give sound, albeit rather cryptic, presentations. The story outlined in Sections 3.3 to 3.7 is given in somewhat greater detail in my own (out of print) text [Arons (1965)]. I have yet to see a college level algebra-based physics textbook that gives what I would regard as a sound, noncircular operational presentation of the Newtonian sequence.

3.8 WEIGHT AND MASS

In the development outlined above, it is to be noted that the term “weight” has never arisen at all, and this should be pointed out, repeatedly, to the students and extracted, in discussion, in their own words. In principle, all the procedures and experiments involved in the operational sequence could be performed in a space ship, away from gravitating bodies, or in a satellite in free fall. Making this explicit helps the students get started on forming the distinction between “weight” and “mass” and fixing the realization that the term “weight of an object” is the name that will be given to a particular force: the *gravitational* force exerted *by* the earth *on* the object, imparting an acceleration of 9.80 m/(s)(s) . This, naturally and directly, becomes the basis for the force arrow, labelled mg , that students will be entering on free-body force diagrams.

In the initial stages, while students are still forming the distinction between the concepts, it is wise to maintain a rigid distinction between the units, speaking of mass only in kilograms and weight only in newtons. Eventually, however, it is impossible to shield students from the looser usage that will be encountered in some technical literature as well as in everyday speech: They will certainly hear locutions such as “a weight of ten kilograms” or “a 3.00 kg weight.” It would be convenient to issue an edict forbidding such usage and wave a magic wand to have this edict enforced, but this will never be achieved (in spite of the most earnest efforts of some purists), and it is better to help the students interpret the inevitable usage as a shorthand reference to the force with which the earth attracts the given body: The phrase “3.00 kg weight” refers to an object on which the earth exerts a gravitational force of $3.00 \times 9.80 = 29.4$ newtons.

Parallel statements would be made, of course, in connection with the British Engineering (BE) system of units. Fortunately, while the country is still inching toward metrification, the majority of textbooks are leading the way by confining themselves to presentation of the SI system, leaving both the BE and cgs systems in abeyance (or placing them in such a way that the instructor can choose to leave them in abeyance). This is the best way to handle the problem pedagogically, not only because SI is preempting the field, but also because throwing all the different systems of units at the students while they are still trying to unravel the concepts is gratuitous. If they need

one of the other systems eventually, they can acquire it at a point where understanding of the basic concepts reduces the matter of units to triviality, and they can close the gap for themselves; it is only before understanding has been acquired that units form a major conceptual obstacle.

Some teachers advocate defining weight as the number measured on a balance or scale, i.e., the force exerted *by* the object *on* the measuring device rather than in the simpler and more direct manner recommended above. Although there is nothing intrinsically “wrong” with this approach [it can be made logically correct and consistent, c.f. French (1995)], I believe it to be unwise and needlessly confusing to students in the initial stages of such subtle concept formation.

Confusion enters because of the inevitable entanglement of the third law—a concept of enormous difficulty for the majority of learners and a hurdle that will be discussed in Section 3.12. The force exerted by the object on the scale changes if one exerts upward or downward forces on the object. The force changes when the system (scale and object) is accelerated upward or downward relative to the earth.

On first encounter it is best to define weight of an object as the the gravitational force exerted by the earth on that object and to show this force directly on free-body diagrams. Concepts and insights have always been acquired and refined by successive approximations, and here is a case in point. Those instructors who wish to expose students to the fact that weight varies from one location to another on the surface of the earth, who wish to emphasize the effect of the earth’s rotation, or who wish to preserve, at all cost, a literal meaning for the term “weightlessness” in free fall, would do better to redefine weight as occasion arises rather than confuse the issue needlessly on the first encounter.

3.9 GRAVITATIONAL VERSUS INERTIAL MASS

We are confronted here with two operationally distinct concepts, yet students have very great difficulty forming the distinction. The difficulty arises partly from the fact that the operational definitions of force and inertial mass are rarely developed with sufficient clarity at the very beginning and partly from the purely linguistic confusion arising from use of the same name for two entirely different ideas.

It is true that one can argue the numerical equality of gravitational and inertial mass from the fact that all objects have the same acceleration in free fall (and this is essentially what Newton does), but this does not provide the student with an adequate *operational* distinction. Furthermore, the student is usually still struggling with the distinction between weight and mass, and invoking free fall at the beginning of the argument simply compounds the confusion. In my own experience, students can be helped to form the distinction by appeal to the following two clearly different gedanken experiments.

Experiment 1: Given the “force meter” operationally developed in Section 3.4 and Fig. 3.4.1, apply it to two different spherically shaped bodies, A and B, and determine their inertial masses through measurement of the accelerations imparted. Suppose we have selected A and B so that their inertial masses turn out to have a ratio of exactly two to one, i.e., $m_A/m_B = 2.00$.

Experiment 2: Now we take bodies A and B and bring them (one at a time) near one of the spheres (body C) at the end of a Cavendish balance. Body C is accelerated by the gravitational attraction, and the Cavendish balance begins to swing.⁴ From the acceleration imparted to body C on the Cavendish balance, we determine the forces exerted on C by bodies A and B (separately) at a fixed distance between centers. We find *by experiment* that the force exerted by A on C is just 2.00 times the force exerted by B on C.

Now we can emphasize the dramatic operational difference between the two experiments. Experiment 1, through the accelerations imparted to the two bodies by the same force, compares the property to which we have given the name “inertial mass.” Experiment 2 has no a priori connection with experiment 1 at all; we are comparing an entirely different property and effect, namely the noncontact forces exerted by A and B, respectively, on a *third* body C. It is truly astonishing that the numerical ratio is exactly the same in both experiments and that this particular order in nature is confirmed experimentally in all circumstances, with all bodies and, by sophisticated indirect measurements, to a fantastically high degree of precision.

How astonishing this is can be dramatized by pointing to the fact that an entirely different interaction between the spheres (say, an electrostatic interaction if they are electrified by rubbing, or a magnetic interaction if they are ferromagnetic and are magnetized) exhibits a ratio of forces exerted on C that bears no relation whatsoever to the ratio of inertial masses of A and B. It is only in the *gravitational* interaction that the ratios are identical.

We now give the property defined operationally by the interaction observed in experiment 2 the name “gravitational mass.” Using the same noun “mass”

⁴The *PSSC* film “Forces” actually shows the execution of a similar experiment: The Cavendish balance consists of a meter stick suspended horizontally at its center from a high ceiling by means of recording tape, which acts as the torsion suspension. Bottles of water hang at the ends of the meter stick. When the balance is stationary (a condition that was achieved only by taking refuge in an isolated, unused building), a box of sand is moved up close to one of the bottles. A spot of light reflected from a small mirror attached to the recording tape provides the optical lever, and the deflection of the spot of light is monitored in the film. This short (10 min) segment dramatically demonstrates the gravitational interaction between ordinary objects and is well worth showing in class if it is available. The only caveat is that the film is purely qualitative, and the objects, not being spherical, do not interact as point masses. With respect to our gedanken experiment, we should eventually be able to argue that our objects interacted as point masses, but this is a refinement that can come later.

for the two entirely different properties constitutes a very unfortunate choice of terminology. It is responsible for much of the conceptual difficulty encountered by the students, but we are stuck with it and cannot change it by fiat. The best procedure is to keep using the adjectives together with the noun and to keep reemphasizing the operational distinction, giving students opportunity to describe it in their own words.

Students gain a clearer picture of the linguistic problem when they see that, regardless of the convention actually adopted, the language might have been quite different. Coulomb, in his great paper on the electrostatic interaction (before crystallization of the term “electrical charge”), refers to the “electrical masses” of his charged spheres. Inverting the analogy, we might just as well have talked about “gravitational charge.” Some other term, neither “mass” nor “charge” would, of course, have been preferable, but we have no choice except to try to make the situation as clear as we can.

Once students have begun to acquire understanding of the preceding operational sequence and appreciate the complete independence of the two experiments, they can come back to the observation that all objects have the same acceleration in free fall and begin to discern the intimate connection among the various observations. Looking at the same idea in more than one way is a powerful aid to understanding the whole scheme, including the distinction between weight and mass.

3.10 UNDERSTANDING THE LAW OF INERTIA

Because of the obvious conceptual importance of the subject matter, the preconceptions students bring with them when starting the study of dynamics, and the difficulties they encounter with the law of inertia and the concept of force, have attracted extensive investigation and generated a substantial literature. A sampling of useful papers, giving far more extensive detail than can be incorporated here, is cited in the bibliography [Champagne, Klopfer, and Anderson (1980); Clement(1989); di Sessa (1989); Gunstone, Champagne, and Klopfer (1981); Halloun and Hestenes (1985); McCloskey, Camarazza, and Green (1980); McCloskey (1983); McDermott (1984); Minstrell (1982); Viennot (1979); White (1983), (1984)].

Learners’ difficulties in encompassing the law of inertia and the concept of force stem in large measure from the wealth of common sense preconceptions and experiential “rules” that most of us assimilate to our view of the behavior of massive bodies before we are introduced to Newtonian physics. Some of these views are Aristotelian (e. g., the necessity of continued application of a push to keep a body moving, it being very difficult to abandon thinking of rest as a condition fundamentally different from that of motion, or to accept the view that, rather than asking what keeps a body moving, we should ask what causes it to stop), but many of these common sense views are more

closely related to the medieval notions of impetus associated with names such as Buridan and Oresme.

All investigations show these “naive” conceptions to be very deeply entrenched and very tenaciously held, and it is important for teachers to understand that student difficulties are not reflections of “stupidity” or recalcitrance. The difficulties are rooted in seemingly logical consequences of perceived order and experience and are vigorously reinforced by insistent use (or actually misuse) of words drawn from everyday speech (inertia, mass, force, momentum, energy, power, resistance) before these words have been given precise operational meaning in physics. Persistent misuse of the terms in thinking to oneself and in communicating with others is a major obstacle to breaking away from the naive preconceptions. (This is another reason for helping the students stand back and become very self-conscious about the process of operational definition—term by term.) Some teachers tend to minimize such problems by labeling them as “merely” a matter of language or semantics, apparently not realizing how formidable and significant the linguistic obstacles tend to be.

Investigations of understanding of the law of inertia further show that it is far from sufficient to inculcate the law verbally and supplement it with a few demonstrations of the behavior of frictionless pucks on a table or gliders on an air track. Many students will memorize and repeat the first law quite correctly in words but, when confronted with the necessity of making predictions and describing what happens in actual physical situations, concretely accessible to them, they revert repeatedly to the naive preconceptions and predictions, giving the disappointed teacher the sinking sensation of not having succeeded in teaching anything at all.

If one wishes to lead the majority, rather than a small minority, of students to understanding the law of inertia, one must accept the necessity of providing a wide array of experiences, both hands-on and hypothetical, in which students make their own errors, encounter the resulting contradictions and, forced by these errors and contradictions, revise their preconceptions. Such experience cannot be provided and mastery developed, however, in one short remedial session. The ideas and initial experiences should be introduced while development of the subject matter is continued without waiting for full mastery on first encounter. One then helps cultivate mastery and understanding through repeated spiralling back to qualitative application of the law of inertia in increasingly rich and sophisticated physical situations as the study of the science continues.

The most effective, albeit fairly expensive, physical situation I have been able to use to such purpose is one in which a full-size 50 lb block of dry ice, with its base smoothed to some degree, is placed on a large glass plate leveled up on a laboratory table. Students are then invited to perform literally “hands-on” experiments (using gloves, of course). A large array of very basic, vitally

important, ideas can be developed Socratically in this context.⁵

- 1 How does the block behave once it is moving? What is the difference between this situation and the one in which ordinary objects slide on ordinary surfaces? (The way in which the block moves in ghostly splendor along the plate, especially at low velocity, without appreciable slowing down, makes a deep impression on most individuals who have never seen such effects.)
- 2 What action on our part is necessary to make the object move faster and faster, that is, accelerate continuously? (To many students it comes as a great surprise that they have to move faster and faster themselves to keep up with the block and to keep on exerting the accelerating force. Even though they previously saw the block move at uniform velocity in the absence of an external force, many of them have not translated this into the sensations that go with the exertion of a constant force on an accelerating object.)
- 3 What is the difference in behavior of the block when acted on by a steady push that keeps up with the block and when it is given a quick shove? (Many students have not had the opportunity to discriminate between a steady force and an impulse. In fact, to many students, the word “force” in the context of setting an object in motion means a quick shove rather than a steady action, and it is important to help them perceive the difference.)
- 4 How large a force is necessary to impart any acceleration at all to the block, that is, is there a threshold effect? (Everyday experience indicates that bodies are not set into motion until a certain minimum force is exerted; this is one of the eminently reasonable, naive rules that students bring with them initially.) The *PSSC* film called “A Million to One,” in which a flea is hitched up and accelerates a massive dry ice puck, is well worth showing if it is available.
- 5 Suppose the block of dry ice is already moving: What must be done to make it slow down very slowly without changing the direction of its motion? Many students are inclined to apply an impulse rather than a gentle, continuous force. They must be guided into doing the latter, and they are usually astonished to find that they must allow their hand to retreat with the moving block. This experience helps reinforce the discrimination between impulse and steady force.

⁵Even though some situation other than the block of dry ice is invoked, the sequence of questions that follows is one through which most students should be led. The difficulties being intercepted are very widely prevalent among students in virtually all introductory physics courses.

- 6 Suppose the block is moving to begin with, and we exert a steady force, either speeding the block up or slowing it down. How does the block behave? Now suppose we make our steady force smaller and smaller. How does the block behave? How will it behave when the force we are exerting reaches zero? (Note that what is deliberately constructed here is a *reversal* of the usual direction of presentation of the ideas: instead of using the zero force situation as the starting point, we are now starting with the nonzero net force and going toward the zero force condition. Many beginning students, at all levels, have very great trouble with the zero force case, despite all the preceding discussion and demonstration. Reversing the line of reasoning and experience, and seeing the situation both ways, helps in the acquisition of the desired insight.)
- 7 Suppose we exert two steady forces on the block in opposite directions, one with each hand. How does the block behave when the one force is larger than the other? When the forces are of equal magnitude?
- 8 Suppose the block is moving: What actions change the *direction* of its motion? (Here, once discrimination between the two has been developed, it is possible to explore the effects of both continuous actions and impulses.) What do you have to do to make the block move at right angles to its initial path? In some other specified direction? In an (approximate) circle? (The principal non-Newtonian expectation found among learners is that an initially moving object will move in the direction of the last impulsive push. It is important that they encounter the contra-intuitive phenomenon personally.)
- 9 What happens if you start the block spinning about a vertical axis? Without using any as yet undefined technical terminology, what are some implications of the observed behavior?

Some words of caution and advice about implementation of this experience: (1) Its essentially personal, hands-on nature tends to reinforce an idea, deeply embedded in many students, that accelerating effects (forces) are necessarily exerted only by animate beings. One should emphasize that contact interactions between inanimate objects (e. g., collisions, release of compressed springs, etc.) also impart acceleration. Noncontact interactions (electric or magnetic) can be introduced or referred to at the teacher's discretion. (2) Although a very small number of students may successfully explore the physical situation without Socratic guidance and emerge, on their own, with most of the insights listed above, the great majority do not carry out a genuine investigation or draw significant inferences under such circumstances. It is essential that the teacher provide guidance, but this is best done by asking questions and eliciting suggestions from the students rather than by giving a set of instructions to be followed. (3) The whole operation is at its best when,

under minimum guidance from the teacher, the students suggest, try, argue, and interpret in their own words, carefully avoiding any, so far undefined, technical vocabulary.

There are, of course, other devices for providing some of the experiences outlined above. A massive dry ice puck does very well on the glass plate, although it is not as dramatic as the 50 lb block. Bricks (or other objects) can be piled on a slab of dry ice instead of using an entire block of the latter. A glass plate is not essential; any very smooth surface will do. A good bit can be done with pucks on an air table, although their rather small mass makes it difficult to perform some of the more delicate experiments, with small forces, using one's own hands. With both kinds of pucks, it is probably better to use some other device for application of a force—a weak rubber band or the stream of air from the hose of a vacuum cleaner operated in reverse, for example.

Another mode allowing for the development of individual experience is, of course, computer simulation, and many groups are developing instructional materials to this end [cf. di Sessa (1982); White (1984)]. Where the tactile, kinesthetic experience with real objects is impracticable, computer simulation is undoubtedly the next best mode. Computer simulation is also useful for providing more extended practice in thinking about a wide variety of examples. It is capable of supplying continual feedback regarding error and correctness and reinforcing the hands-on observations after the latter have been carried out. The weakest mode is that of lecture demonstration in which student participation is passive—limited to hearing assertions and to seeing effects produced by someone else.

Pencil-and-paper questions and exercises are also a useful component of instruction. They can be designed to help the student confront contradictions in his or her own thinking and to converge on genuine grasp and understanding. Such questions play an especially important role in homework and on tests and examinations; the appendix to this chapter contains selected examples.

3.11 WHAT WE SAY CAN HURT US: SOME LINGUISTIC PROBLEMS

There are natural tendencies in everyday speech that are inimical to development of understanding of the concept of force and the law of inertia. Teachers should become sensitive to these usages, learn to avoid them themselves, and divert students from their use. Some examples:

- 1 There is a very strong, almost universal, tendency to say that a force (or a net force) causes a body to “move.” Students should be led to say “accelerate” instead of “move.” The word “move” seriously obscures the issue and tends to sustain an Aristotelian view. Students who use it tend to fix on its connotation of “velocity” and lose sight of the primacy of

“acceleration,” particularly in the early stages when acceleration is still an unfamiliar concept and is incompletely distinguished from velocity.

- 2 A very common locution is that “force overcomes the inertia of a body.” This encourages the student in thinking of inertia as a force to be “overcome” by other forces. (It is true that Newton himself listed “vis inertiae” as one of the forces to be discerned in nature, but he avoided confusing this with “motive forces” that impart changes in momentum.) In modern instruction, it is best to avoid any implication whatsoever that inertia is a kind of force.
- 3 “Force” is interpreted by many students as something *given* to, being a *property* of, or *resident* in a moving body or one being accelerated. (How much is this reinforced by our tendency to talk about forces “imparted” to a body? I myself find the latter locution difficult to avoid.) In any case, it is advisable to counter this notion and to emphasize external effect and interaction, as opposed to residence in the body.
- 4 The meaning of “net,” “resultant,” or “total” force (when forces are acting simultaneously on a given body) should be developed very carefully and explicitly. There is a strong tendency among students to think of some of the individual forces as having disappeared, or having been somehow obliterated, in the superposition, especially when some of the forces oppose each other and are “overcome” in the final effect. Some students, when one force “overcomes” an opposite force, see the dominant effect as acting alone, not as the algebraic or vector sum of the two.
- 5 Confusion between a continuous action and an impulsive shove in connection with “exerting a force” has been mentioned in the preceding section. The language requires explicit attention.
- 6 Many students proceed to talk about forces as “working” on objects when dynamic situations are being considered. It is advisable to intercept this locution and stick to the word “acting.” Casual use of the word “working” invites confusion when one builds the energy concepts later.

3.12 THE THIRD LAW AND FREE-BODY DIAGRAMS

The third law is, of course, part of the auxiliary “text” essential for full understanding of the concept of force. Without it there is no basis for separating two or more interacting objects and applying the second law to one object at a time, and, without it, students are seriously delayed in developing a comprehension of what object does what to which in familiar physical interactions. Those authors who develop the second law and then proceed to conservation

of momentum as though that takes care of all the necessary physics, leave their students crippled through inadequate understanding of the force concept.

Once we are used to it, the idea articulated in the third law seems transparently simple, and teachers tend to become insensitive to the very great difficulty the majority of students encounter. Difficulties arise in a number of sources and become compounded for many students.

- 1 *Forces exerted by inanimate or “rigid” objects.* As pointed out in Section 3.10, many students have the preconception that forces can be exerted only by living beings, and they balk at the idea of a table, a floor, a block exerting a force on anything. As a college student once said to me in exasperation, “How can the table exert a force on the book? It has no p-p-power!” Thus, even though they see the table as a “barrier” to downward motion of the book, many students do not see it as exerting an upward force. Similarly, they do not see “resistance to movement” from a surrounding fluid medium, or from rubbing at surfaces, as a force.

This is *not* a trivial conceptual problem, and, since very few texts provide explicit help, it is up to the teacher to develop the insight. Most students are willing to accept the idea that deformed objects (e. g., springs) that return to their initial configuration are capable of exerting a force, and this provides an effective starting point. Because they are aware of the deformation, they can be led to admit that the bed, sofa, easy chair exert an upward force on the sitter, but they regard apparently “rigid” objects as being qualitatively different and do not readily visualize decreasing, but nonzero, deformation as rigidity increases. It is quite difficult to convey the realization that the table, floor, and block also deform—even when loaded with a sheet of paper. Minstrell (1982) describes how he finally convinced a group of students that the laboratory table deforms when loaded: He directed the beam of an overhead projector so that it was obliquely reflected from the surface of the table to an adjacent wall, thus making an optical lever. When the students saw the spot on the wall being displaced as a student walked on the table, they began to accept deformation of apparently rigid objects.⁶

⁶Students need explicit help and guidance in learning to visualize effects that elude direct sense perception. The deformation of apparently rigid objects in the context now under consideration is usually the first opportunity in a physics course, and its importance should not be underestimated. Later, such visualization is essential to understanding what happens in elastic and inelastic collisions, in deformations under tension and compression, in the breaking of a string, in the rupture of a container of water when the water freezes, in the propagation of longitudinal and shear waves in solids, in understanding that the far end of a long steel rod is not displaced at the same instant we push on the near end, and ultimately, to being prepared to accept finite time intervals for the transmission of electromagnetic effects (i. e., the invention of field theory). The sequence of visualization and concept building is best initiated at this, seemingly trivial but nonetheless crucial, starting point.

- 2 ***“Passive” versus “active” forces.*** In light of the difficulties cited in item 1, it turns out to be helpful for students to distinguish between two classes of forces, designated as “active” and “passive,” respectively. Active forces are exemplified by animate pushes and pulls, the gravitational force, electric and magnetic forces. Passive forces are defined as those that arise, and adjust themselves, in response to active ones, for example, in compression of a spring, deformation of the table or floor under the load of a block, frictional forces, and so on. The increase (or “adjustment”) of the passive force cannot take place indefinitely; it continues only to the point at which something breaks (table or floor or string) or gives way (as in sliding friction).
- 3 ***Stating the third law.*** The old, conventional jargon “for every action there is an equal and opposite reaction” has always been gibberish to the majority of students and, fortunately, many authors are abandoning it. It is best to say “if one object exerts a force on a second, the second exerts an equal and opposite force on the first”—or some other, equally simple and straightforward, form. Even this simple a statement is not initially understood. Students, even when repeating the words correctly, do not do so with the clear realization that one is talking about *two* different forces, each acting on a *different* body. They need extended help in building this realization and making it explicit in diagrams and in their own words.
- 4 ***Noncontact forces.*** Confusion concerning the simultaneous presence of two different forces acting on different objects is enhanced by the fact that, at these early stages of development, we tend to concentrate almost exclusively on *contact* forces and, in the case of contact forces, it is difficult to discern the two separate actions. Also, in the case of the only noncontact force usually considered (namely gravity), we postulate the interaction on the basis of the observed acceleration of free fall, and we are unable to demonstrate the force, equal and opposite to the weight of the object, that is exerted by the object on the earth. To most students this second force remains a source of mystery, confusion, and, in large measure, disbelief.

Without going into details about static electricity or magnetism, it is very helpful at this stage to invoke these effects simply to the extent of demonstrating noncontact interactions made evident by the observed accelerations. Two charged pith balls visibly attract or repel each other without contact; thus, we are forced to conclude that each experiences a separate force. Two bar magnets attract or repel each other without contact. Two air track gliders, with appropriately mounted magnets, undergo collisions without making contact. (This effect startles many students.)

A charged rod held in our hand attracts or repels a suspended pith ball or visibly accelerates bits of paper lying on the table. After discussion of the earlier demonstrations, it becomes plausible to the students that not only the pith ball and the bits of paper but also the rod experiences a force, even though the latter force eludes our physical sensation. The same applies to the case in which the magnet, held in our hand, accelerates small nails. With sufficiently strong magnets and more massive objects, the noncontact interaction can be sensed directly.

Given these demonstrations, the third law becomes much more plausible and intelligible to many students. Their force diagrams improve, and the gravitational force exerted by the book on the earth is accepted as reasonable and consistent, however undetectable it might be.

5 *Drawing free-body diagrams.* It is a well-known phenomenon that many students, when they first start drawing free-body force diagrams, produce pictures resembling a porcupine shot by an Indian hunting party—pointed entities stick out randomly in all directions. Practice in analyzing familiar, everyday situations is essential. As the randomness diminishes, many students still persist in showing the two equal and opposite forces of the third law acting on the same body. To at least some extent, these tendencies are fostered by many textbooks: A block is shown resting on the floor and, to save space, the two interacting objects (block and floor) are not shown *separated*. The force exerted by the floor on the block and the force exerted by the block on the floor thus appear on the same picture instead of on well separated pictures, and the message about two different forces acting on different objects is completely obscured. Furthermore, the two forces are rarely described verbally right on the diagram itself.

Lecturing to students about these problems, telling them what should be done, and drawing diagrams for them produces very little effect. A more effective procedure is one that requires students to construct diagrams of their own (including redrawing the faulty diagrams in the textbook) under the following rules:

- (a) *Both* objects in each relevant interaction should be shown: In the case of the book resting on the table, both the book and the surface of the table should be shown in well separated diagrams, even if the book is the principal focus of attention. The third law pair of forces between the book and the surface of the table should be shown, each on its appropriate diagram. In the early stages of such exercise, the earth should be shown as well, since it is the other object involved with the gravitational force acting on the book. (As time goes by, and the majority of students absorb the idea that, in the case of the weight of an object, the other member of the third law pair

is visualized as acting at the center of the earth, one can begin to drop the requirement of including the earth.) When objects are connected by strings, there should always be a well separated force diagram of the connecting strings as well as of the other objects, even when the strings are regarded as “massless.”

- (b) Every force should be described in words right along with the diagram. A verbal description means indicating the nature of the force and stating what object exerts the given force on what, for example, gravitational force exerted by the earth on the book; gravitational force exerted by the book on the earth; normal contact force exerted by the book on the table; frictional force exerted by the table on the book; contact force exerted by the string on body A; contact force exerted by body B on the string; and so on.
- (c) After the arrows are drawn and then described in words, each third law pair should be identified explicitly.

It is the *combination* of being aware of active, passive, contact, and non-contact forces, drawing arrows on well separated pictures, describing the forces in words, identifying third law pairs, and being corrected on their errors, that gradually leads students to understanding of the third law and the ability to set up problems and to apply the second law without guesswork and memorization. As in all other instances involving subtle concept formation, the practice must be spread out over time; attempts at quick remediation invariably fail.

3.13 LOGICAL STATUS OF THE THIRD LAW

In the *Principia* Newton felt it necessary to justify Law III, and he does this in the lengthy Scholium that follows the enunciation of the three Laws of Motion. First he cites papers that Wallis, Wren, and Huygens had (separately) contributed to the Royal Society in 1669 in which they each cited conservation of “quantity of motion” (momentum) in “impact” (collisions) as a fundamental law of motion [see Arons and Bork (1964)]. He then argues that such conservation follows from the third law and even implies that Wallis, Wren, and Huygens obtained their insights by having *used* the third law (something that is quite unlikely, since conservation of momentum in collisions had been recognized empirically for some time without clear articulation of a force concept.)

He cites colliding pendulum experiments of his own as providing corroborative evidence for momentum conservation and goes on to present the following argument appealing to “attractions,” which at that time was the technical term for the (noncontact) electrostatic and magnetic interactions:

In attractions, I briefly demonstrate the thing after this manner. Suppose an obstacle is interposed to hinder the meeting of any two bodies A and B, attracting one the other: then if either body, as

A, is more attracted towards the other body B, than the other body B is towards the first body A, the obstacle will be more strongly urged by the pressure of the body A than by the pressure of the body B, and therefore will not remain in equilibrium: but the stronger pressure will prevail, and will make the system of the two bodies, together with the obstacle, to move directly towards the parts on which B lies; and in free spaces, to go forwards in infinitum with a motion continually accelerated; which is absurd and contrary to the First Law. . . I made the experiment on the loadstone and iron. If these, placed in proper vessels, are made to float by one another in standing water, neither of them will propel the other; but, by being equally attracted, they will sustain each other's pressure, and rest at last in equilibrium. [Note that Newton speaks of using a "loadstone and iron," not two loadstones, i.e., in his experiment one of the objects is passive.]

It is very helpful to the students to invoke this example since it greatly expands and enriches the initial context in which the third law is usually presented. An analogous experiment is also easily performed with gliders on an air track.

Newtonian theory is frequently referred to as an "action at a distance" theory, and the third law lies at the heart of this description. The third law says that *all* interacting objects exert equal and opposite forces on each other *instant by instant*, and this applies to widely separated gravitating bodies as well as to bodies exerting contact forces on each other: Zero time elapses between a change occurring at one body and the effect of the change being felt at the other.

If we push on one end of a long rod, the other end of which is in contact with a block, the block does *not* exert an equal and opposite force on the rod at the same instant we push. A finite time interval elapses between our push and an effect at the block, the time interval being determined by the velocity of the elastic wave that passes down the rod. Thus, Newton's third law does *not* hold, instant by instant, for the forces at either end of the rod; it holds only layer by layer of material along the length of the rod, and momentum and energy are both conserved only by virtue of propagation of the elastic wave.

Throughout the later years of his life, Faraday was deeply concerned with analogous situations in electricity and magnetism: If two electrically charged particles are at rest, exerting equal and opposite forces on each other, and one of the particles is suddenly displaced, changing the force to which it is being subjected, does a time interval elapse before the force on the other particle changes? Does the compass in the Oersted experiment begin its swing at the instant the current is initiated in the wire or does a finite time interval elapse? He constructed delicate mechanical equipment designed to detect such time intervals but, of course, never succeeded.

Maxwell appreciated the significance of these questions, and his invention of the first field theory provided an answer as well as a model for all subsequent field theories.

The point is that the third law does *not* always hold, and this is why modern physics has given primacy to conservation of momentum in the hierarchy of physical law. Although one would not discuss all these aspects with students at the time of first introduction of the third law, it is well to start laying the groundwork for eventual perception of where the law fails. The rod pushing the block makes a good starting point. The students are initially completely incredulous concerning the finite time interval, and the incredulity can be shaken by pulling on the block with a long slinky. One can spiral back to these questions, and fill in gaps, on arriving at discussions of mechanical wave phenomena and at the appropriate points in electricity and magnetism.

3.14 DISTRIBUTED FORCES

Very few textbooks lead the student to perceive that the single arrows representing the weight of an object, or the normal force on the object at an interface, or the frictional force at the interface, are a shorthand for the sum of distributed effects that must be added “chunk by chunk.” This idea is usually left to implication, and only a very few students perceive the implication. Although naive students do not articulate the idea explicitly, they tend to hold the unexamined view that the arrows represent concentrated effects akin to actions such as pushing with a finger or pulling on a string. Later on, the lack of comprehension of distributed effects seriously impedes their understanding of the origin of buoyant forces acting on bodies in a fluid or hydrostatic pressure in general.

Summing the distributed effect does not seem to be an especially difficult idea for students to absorb once it is called to their attention. The point is that it *does* have to be called to their attention. If this is not done, a conceptual gap remains, and many students do not close this gap spontaneously until very much later in their development.

3.15 USE OF ARROWS TO REPRESENT FORCE, VELOCITY, AND ACCELERATION

While one is confined to a single context (forces alone, velocities alone, etc.), the use of the arrow symbol to represent the given quantity causes no confusion. When we start dealing with situations in which forces are imparting acceleration to a body having nonzero velocity, however, use of arrows of identical form to represent all three different quantities does cause confusion in many students. They interpret velocity and acceleration arrows as forces acting on the body, and, in drawing their own force diagrams, they gratuitously

insert velocity and acceleration arrows as additional forces. (Such confusion arises, for example, when one wishes to examine all the effects on an object in projectile motion; when one deals with objects in an accelerating car; or when one is concerned with forces applied to, and the velocity and acceleration of, a bob in circular motion.)

This confusion can be countered to some degree by slightly altering the notation. My own system is to use the ordinary arrow for force, a single-half-headed arrow for velocity, and a double-half-headed arrow for acceleration as in Fig. 3.15.1.

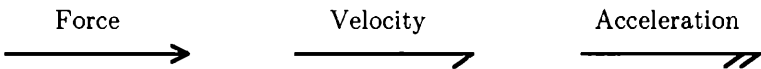


Figure 3.15.1 Using different arrows for different vector quantities.

I ask the students to use this notation on tests and homework, and I use it myself in lecture presentations (as I use it throughout this book.) The system is not onerous, and it helps reduce the inclusion of velocities and accelerations as forces on free-body diagrams. There is, of course, nothing sacred about this particular notation, and any other form (such as color coding) that distinguishes the quantities would serve equally well.

3.16 UNDERSTANDING TERRESTRIAL GRAVITATIONAL EFFECTS

Interviews with students reveal extensive misconceptions and confusion about “gravity” and gravitational effects—misconceptions that are rarely spontaneously articulated by the students, that frequently pass unnoticed by teachers, and that seriously impede understanding of the material being taught.

- 1 *Meaning of the word “gravity.”* One semantic problem, originating in early years and persisting to college level in many students, stems from an answer provided by many teachers and parents when the child asks, “Why do things fall?” A very common answer is, “*Because of gravity.*” (If you ask this question of a class of college students, you will get the indicated answer in the majority of cases. Only a few students are uneasy about such a facile answer and fewer still have the self-confidence to challenge it in the way it should be challenged.)

Children, as well as many adults, take this answer very literally: since the word “because” has been used, they uncritically jump to the conclusion that a *reason* has been given—that the “why” has been answered. They naively believe that a scientific name provides a reason; much of their experience with science in the schools has reinforced this acquiescence.

Students should be made aware of some of the history of the term: The Greeks endowed bodies with the teleological properties of “gravity” and “levity,” representing built-in desires or tendencies of the bodies to seek the center of the earth or to rise toward the celestial domain; 17th century science eliminated both the teleology and the term “levity” and applied the name “gravity” to the observed interaction between objects and the earth. With the Newtonian synthesis, the meaning is expanded by the grand perception that the same effect that makes the apple fall also binds the moon to the earth and the earth and planets to the sun, eventually encompassing all material objects.

Finally, however, students must be made explicitly aware that the name does nothing more than conceal ignorance—that to this day, and despite the power of the Newtonian synthesis and the beauty of the general theory of relativity, we have no mechanism for the interaction and no idea of how it “works.”⁷ It is interesting to note what Galileo had to say about this matter. In the *Dialogue Concerning the Two Chief World Systems* one finds the following exchange:

SIMPLICIO: The cause of this effect [what it is that moves earthly things downward] is well known; everybody is aware that it is gravity.

SALVIATI: You are wrong, Simplicio; what you ought to say is that everyone knows that it is called “gravity.” What I am asking you for is not the name of the thing, but its essence, of which essence you know not a bit more than you know about the essence of whatever moves the stars around. I accept the name which has been attached to it and which has been made a familiar household word by the continual experience we have of it daily. But we do not really understand what principle or what force it is that moves stones downward. . .

It seems that the appropriate form of the dialogue has not changed very much over the interval of almost four hundred years.

Helping students see that names, as such, do not constitute knowledge or understanding, and coupling this with the emphasis on careful operational definition advocated throughout this book, does much to put

⁷In using the words “mechanism” and “works” I am referring to processes that we visualize in terms of ordinary sense experience. We visualize such microscopic effects as gas pressure and diffusion, evaporation and condensation of liquids, crystallization and structure of solids, in terms of familiar behavior of macroscopic particles. We visualize invisible elastic waves in solids in terms of what we have seen happening on soft springs. We visualize the propagation of classical electromagnetic waves in terms of an analogy to mechanical shear waves. We have no corresponding forms of visualization for quantum mechanical effects or for gravitational interaction, “virtual” entities notwithstanding.

students in the position of recognizing when they do not know the meaning of a technical term and to recognize when meaning has, or has not, been provided. My own observations show that many students cease name-dropping of terms they have picked up but do not understand, and many report asking for meaning of technical terms in other (not necessarily science) courses.

- 2 *Meaning of “vertical” and “horizontal.”* Very few students possess clear operational definitions of “horizontal” and “vertical.” If asked how they might, as simply as possible, establish a precisely vertical direction right where they happen to be, many respond, “perpendicular to the ground.” If one suggests going over to the steep slope of a nearby hill and establishing the perpendicular to the ground, they back away from the initial suggestion, but few have anything with which to replace it. All told, very few students have established a clear connection between the direction of the force of gravity and the meaning of “horizontal” and “vertical”—either via the plumb bob or the carpenter’s level.
- 3 *Air and gravity.* Many students, especially among the nonscience oriented, acquire the information that the air (or the atmosphere) “presses down on things” and translate this into an association with gravity. They thus tend to view gravity as imposing a downward push rather than a downward pull: Air presses down on the book on the table; gravity “disappears” when air is removed; many expect that objects would float around in an evacuated bell jar without the air to hold them down. Very large numbers of students expect an air-filled balloon, which is seen not to float in air, to float in an evacuated bell jar. One should allow these expectations to be openly articulated and brought to the surface, and one should then counter them with suitable demonstration experiments.
- 4 *Meaning of “vacuum.”* A concomitant difficulty arises with the word “vacuum.” Once in conducting a discussion of some observations of naked eye astronomy with a class of preservice elementary school teachers, I casually referred to the “vacuum of outer space.” Noticing strange looks and sidelong glances among the students, I pursued the issue and finally discovered that, where I was thinking of space devoid of matter, most of the members of the class were thinking of the household appliance they used for cleaning rugs. They were left wondering what motivated me to talk about some mysterious cosmological vacuum cleaner. I forthwith brought out a pump, a hose, and a bell jar.
- 5 *Uses of the feather and coin tube.* The classical demonstration of the “feather and coin” tube (in which objects that clearly do not fall together in air do so in a vacuum) is well worth showing in virtually all classes. (The only students likely to have seen it are those who happen to

have had an unusually good high school physics course.) Not only does this apparatus demonstrate Galileo's law of falling bodies, but it also offers the opportunity to raise the issues of "vertical" and "horizontal," discuss the meaning of the word "vacuum" and, for those who expect gravity to disappear in the absence of air, emphasize that this is not what happens. The films taken of this phenomenon by astronauts on the moon are well worth showing for their broad range of interest, but they are no substitute for seeing the effect in the tube evacuated right there in class.

- 6 *Meaning of g .* A very large number of students, including those in calculus-based physics courses, when asked what the symbol g stands for in kinematics and dynamics, respond "gravity." They do *not* invoke the word "acceleration" at all. When the questioning is pursued further, it almost invariably emerges that students who respond this way have no understanding of any of the things they do with this symbol and are simply trying to memorize problem-solving procedures. They cannot clearly identify the kind of quantity the symbol represents, although many students seem to regard it as being more a force than anything else. It is necessary to get these students to the point at which they give a correct interpretation of g in their own words, with physical illustrations of its meaning in everyday experience.
- 7 *"Feeling" the weight of an object.* Teachers and textbooks frequently say that "we feel the weight of an object when we hold it" and imply that the same force acts on the table supporting the object. Granted that we can get away with this locution in everyday speech, it can be very damaging, however, in a physics course in which we should be trying to ensure precise understanding of the scientific concepts and language we are creating.

The above locution is valid *only* if "weight" has been defined as the force exerted by the object on the measuring balance or scale. If, as is greatly to be preferred, "weight" has been defined as the gravitational force exerted by the earth on the object, the locution is *not* valid.

As discussed in Sect. 3.8, the term "weight of an object" is best introduced, and then reserved *exclusively* for the *gravitational* force exerted *by* the earth *on* the object. Given this meaning, the force we feel when we hold an object is *not* the weight of the object but the contact force the object exerts on us. It is true that this contact force is sometimes numerically *equal* to the weight of the object, but the equality does not make it the same force. In fact, the two forces are not even numerically equal if something is pressing down or tugging upward on the object or if we are accelerating the object up or down. The distinction between the two forces is not trivial and, if it is not maintained, a

large measure of understanding of the scientific vocabulary is lost. Furthermore, understanding of the third law pair at the interface (the force exerted by the object on the table and the force exerted by the table on the object) is undermined.

- 8 *Weight and weightlessness.* Most teachers are aware of the unfortunate use (or misuse) of the word “weightlessness” in connection with satellites and space vehicles. There is not much we can do about the usage (any more than we shall be able to force people to say “mass” instead of “weight” when talking about a number of kilograms of potatoes in a grocery store). We can, however, give students an understanding of what is being described and why the terminology is unfortunate.

Some authors and teachers try to dodge the issue by suddenly switching the meaning of the word “weight” (usually without openly confessing that a switch is being made): After initially defining “weight” as the gravitational force exerted on the object by the earth, they switch to describing “weight” as the reading on the platform scale on which the object is supported, that is, they transfer the designation to apply to the normal force exerted by the object on the platform. As pointed out in the preceding paragraphs, not only is this *not* the gravitational force exerted by the earth on the object but, in many circumstances, it is not even numerically *equal* to the gravitational force. Although this usage may *seem* to simplify matters for the learner, it is invariably disastrous and plants far more difficulty and confusion than it mitigates. As an illustration of such confusion: This usage reinforces the mistaken notion that the force of gravity indeed vanishes when an object is in free fall or when it is removed to appreciable distances from the earth.

The best procedure is to stick unswervingly to the initial definition of “weight” as the gravitational force acting on the object and help the student analyze the sensations he or she personally experiences under various circumstances:

First one must lead the student to realize that we do not sense or feel the gravitational force itself; we postulate its existence on the basis of the observation of acceleration in free fall and the definition of “force” as an action that imparts acceleration. When we jump from an elevated position, we do not feel something tugging on us as we are falling.

Next we lead the student to recognize that what we *do* sense or feel is the normal force exerted on us by the object we stand or sit on. This force is numerically equal to our own weight only if no one is sitting on our shoulders or trying to lift us, and only if we are not being accelerated either up or down. (Thus the student can be led to define the very special circumstances under which we “weigh ourselves.”)

Now we proceed to explore what happens to the reading on the platform scale as we are accelerated up or down—say, in an elevator.

Most students have noted the sensations that go with such accelerations and are prepared for interpretation of the forces they experience: An upward force larger than the one normally felt when the acceleration is upward; an upward force smaller than the one normally felt when the acceleration is downward.

Finally one can argue to the limit: What happens to the upward force exerted on us by the platform as the downward acceleration gets closer and closer to that of free fall? Most students readily agree that the upward force on us, and the reading on the scale, go to zero.

One can now take up the matter of terminology: When we are in free fall, the gravitational force exerted on us by the earth has *not* become zero. What has become zero is the normal force at our feet—the force that we do sense directly. Under these circumstances we experience a strange sensation, one that might be called a “sensation of weightlessness.” Hence arises the poor terminology in which the word “weightlessness” is used to describe the situation in a freely falling elevator or in a satellite. We must understand the confusing usage and not interpret the word as literally meaning that the gravitational forces have become zero.

9 Forces in free fall and in projectile motion. Many authors and teachers have become so accustomed to Galileo’s law of free fall and to the usual idealizations (“thinking away” the ever-present frictional effects) that they are tempted to traverse this subject matter as quickly as possible in order to extend coverage to more “interesting” things. Unfortunately, the common sense preconceptions pervading this area are very tenacious, and many students, if not given the necessary help, are left so far behind that they take refuge in memorizing and never really catch up.

- (a) Many students, when they finally open up, tell me that they were “told,” and that they can readily repeat the statement, that all objects fall together when dropped, but they have “never really believed it.” They need to see and *discuss in their own words*: simple demonstrations such as the dropping of a sheet of paper side by side with a similar sheet crumpled up into a ball; the dropping of the sheet of paper placed on top of a falling book; stroboscopic pictures of large and small objects falling side by side; the feather and coin tube mentioned above, etc.
- (b) After becoming convinced that all objects do indeed fall together in the absence of rubbing effects, many students will then switch to the view that, in order for this to happen, the forces acting on the different objects must all be the same. Countering this requires discussion and observation; a simple assertion on the part of the

teacher produces little effect.

- (c) Students should have the opportunity (in homework and on tests) to draw their own force diagrams (including both the object and the earth) for: An object dropped from rest; an object thrown vertically upward (on the way up, on the way down, and at the top of the flight); a frictionless puck sliding along an air table and then the same frictionless puck while flying through the air after having sailed off the table; a projectile at various points in its trajectory.
- (d) The force diagrams in (c) should, in each case, be accompanied by a *separate* diagram showing the instantaneous velocity vector and by still another diagram showing the instantaneous acceleration vector. The juxtaposition of these various diagrams is significant in enhancing understanding since it makes the student view the same situation in entirely different ways.

10 *Student views surprising to many teachers.* Gunstone and White (1981) present a highly revealing set of student responses concerning the following situation: A bicycle wheel is mounted as a pulley with its axis 2 m above the laboratory bench. A cord, connecting a bucket of sand and a block of wood, equal in mass, is placed over the pulley, that is, the students see an Atwood machine with a bucket of sand at one end and a block of wood at the other. (The students participating in the investigation were first-year students at Monash University in Melbourne, Australia—students who had not yet had university instruction in physics.) The students were then asked various questions, including ones that required making predictions as to what would happen when certain changes were made, and they were asked to write out the reasons for their answers.

- (a) The participants were shown that the pulley rotated freely, and then the cord was placed over the pulley in such a way that the bucket was markedly higher than the block. The system remained stationary. The participants were asked, “How does the weight of the bucket compare with the weight of the block?” Of the participants, 27% said the block was heavier, the largest proportion of these explaining their conclusion by pointing to the fact that the block was nearer to the floor and thus must be heavier. Another reason given by some students was to the effect that “Tension exists at both ends of the string. At the end towards the bucket the tension is less than at the end towards the block. This then causes the block to pull itself down and thereby raises the bucket.”
- (b) The students were then asked to predict what would happen if a large scoop of sand were added to the bucket. Now 30% predicted

that the system would shift to a new equilibrium position with the bucket closer to the table and the block higher up.

- (c) After it was shown that the system moved continuously after the scoop of sand was added to the bucket, the participants were asked to predict how the speeds of the bucket would compare at two marks—one high and one low (near the table). Although 90% correctly predicted that the speed would be higher at the low mark, some indicated that their prediction was based on knowledge that the gravitational force acting on the bucket increased as the bucket went down (or the force on the block decreased as it rose). Others stated that the acceleration of the bucket would be g . When the demonstration was made, 7% of the students reported observing the speeds to be equal at the two marks. The reconciliations of prediction and observation among these students included “no net force,” “objects only accelerate in free fall,” “friction,” and “error in observation.”
- (d) The block and bucket (equal masses) were placed on the pulley so that they hung at the same level without motion. The block was then pulled down about 0.7 m and held. Students were asked to predict what would happen when the block was released. Only 54% predicted the system would remain stationary; 9% predicted return to the original position; 9% predicted the bucket would fall; 2% predicted the block would fall.

Gunstone and White give many more details in their informative paper; the preceding highlights have been selected for illustration. The moral of these illustrations is that we, as teachers, become so familiar with these basic concepts and phenomena that we regard them as too trivial to command any time in instruction. Only when questions of this variety are included in both homework and tests, however, do we begin to help the large number of students who have such difficulties achieve understanding.

3.17 STRINGS AND TENSION

Many textbooks bring forth the word “tension” and start using it as though everyone must know what it means without operational definition. The student is confronted with the familiar problem in which one string is stretched by opposite forces of 50 N at each end while a second string, with one end attached to a wall, is pulled with a force of 50 N. The student wonders how it is possible for the tension in the string to be the same in each case and is unable to see why it is not 100 N in the first string.

There are two difficulties superposed here. One is that, when this situation is first encountered, many students have not fully assimilated the third law

and, not drawing an adequate force diagram of the string, fail to see that the two situations are identical as far as the forces on the strings are concerned. The other difficulty, however, is that “tension” has not been defined.

One simple approach is to lead the student to imagine “cutting” a stretched string at some point along its length and drawing the forces acting on the two segments. (Not only is this a good exercise in using the third law, but it also introduces students to the examination of forces in the *interior* of objects. Up to this point all forces and force diagrams have usually been confined to external effects, and the realization has not been formed that one can, in imagination, “cut through” an object and show the forces at the selected cut.) Having drawn the equal and opposite forces acting on the two segments at the cut, one can give the name “tension at the cut or section” to the magnitude of the force acting on either segment. Tension and compression in rods or columns can then be defined in a similar way.

Having defined tension in this way, it is now a relatively simple matter, inviting valuable phenomenological thinking and visualization, to examine the tension in a massive rope (or chain or rod) as the object is accelerated by a force at one end. It is not necessary to solve quantitative problems! As one examines the tension “chunk-by-chunk” through the length of the object, it becomes apparent, through application of the second law, that it must decrease continuously from a value equal to that of the applied force at one end to zero at the other. One can then leave for homework the further problem of how the tension varies when a rope is accelerated with two opposing forces, unequal in magnitude, at each end.

3.18 “MASSLESS” STRINGS

It is well known that “massless” strings are a source of significant conceptual trouble for many students. They have no intelligible operational definition of “massless”; they fail to see why the forces of tension should have equal magnitude at either end; they proceed to memorize problem-solving procedures without understanding what they are doing. The principal problem here is that, when massless strings come along in the text, many students have not yet fully assimilated the idea that the difference between the magnitudes of oppositely directed forces acting on an accelerated body depends on the mass of the body (when acceleration is fixed). A careful and clear development of the “massless string” concept therefore not only helps students in this immediate kind of problem solving; it helps students register a vital aspect of the second law that has so far eluded them.

Understanding the meaning of “massless” is greatly facilitated by leading students through an operational definition of tension (as in the preceding section) and then proceeding with something like the following sequence: Suppose a rope of mass m_R is attached to a massive block, and we accelerate the system horizontally, with an acceleration a_R , by pulling on the end of the rope with a

force T_1 . Separate free-body diagrams of the rope and the block should then be drawn, and students can be led to acknowledge that the block exerts a force on the rope at the opposite end; denote this force by T_2 . Further discussion is usually required to make sure that students understand that T_2 must be smaller in magnitude than T_1 and that these two forces are also equal to the (different) tensions at the two *ends* of the rope.

Now they apply the second law to the rope, obtaining the expression:

$$T_1 - T_2 = m_R a_R \quad (3.18.1)$$

Students must be led to interpret this expression. At this stage of development, very few students understand what it means to interpret an algebraic expression, and there is massive resistance to doing so. They should be led to say that the equation indicates that the two forces are equal in magnitude when the acceleration is zero and that the equation confirms the earlier, qualitative, conclusion that T_1 is larger than T_2 when the acceleration is not zero.

Now it is possible to get at the real point at issue: What happens to T_2 as the mass of the rope m_R is made smaller and smaller while the acceleration a_R is kept fixed? Having reached this point, most students are able to discern that T_2 becomes more and more nearly equal to T_1 , that, in the limit, the two tensions are equal, and that this is the real meaning of the concept of “massless string” in the context of the textbook problems.

A somewhat more rigorous development, highly desirable for more sophisticated students, is to set up the algebra for the entire system (including the block, with mass m_B , ending with the expression

$$\frac{T_1}{T_2} = 1 + \frac{m_R}{m_B} \quad (3.18.2)$$

Interpreting, this equation shows that T_2 becomes very nearly equal to T_1 when m_R is very small compared to m_B , and shows the students that “masslessness” is, in the final analysis, a relative and not an absolute matter.

Such an analysis gives students in engineering-physics courses, for example, a very rudimentary exposure to theoretical formalism—an exposure that is, unfortunately, denied them in most instances through neglect of available opportunity. Teachers then wonder why the students seem to be so naive on such matters in more advanced courses.

3.19 THE “NORMAL” FORCE AT AN INTERFACE

The normal force N is usually first encountered in situations in which an object of mass m rests on a horizontal surface: the book on the table, the student’s own body on the ground. In this special case the normal forces exerted by the book on the table and the table on the book happen to be equal in magnitude

to mg , the weight of the book. Many students, not yet having formed a clear understanding of the force concept and of the third law, simply memorize the statement $N = mg$ more or less in self-defense and continue to stick to this equation in circumstances in which it is not applicable. [Locutions about “feeling the weight of the object when we hold it up,” discussed in Section 3.16 (7), also feed this misconception.]

To forestall this difficulty, students should be led to visualize how the normal force varies when they exert an upward tug on the book and when they press down on it vertically, and this should be done as soon as possible after they have begun to accept the idea that the inanimate table is indeed capable of exerting such a force. They should be led to articulate the insight that, in fact, N is almost never equal to mg , and that the equality obtains only in the very special case in which there are no other vertical forces acting besides the pull of the earth.

Another exercise that is very helpful at this point, repeating some ideas but altering and enriching the context, is to press the book against the wall. Now the wall, another inanimate object, must be conceded as capable of exerting a normal force, and this normal force has nothing at all to do with mg ; its magnitude is determined exclusively by the horizontal force we exert with our hand. (This situation is also useful for showing students that frictional forces do not necessarily depend on mg , a misconception they also pick up from the first encounter with friction on horizontal surfaces. See Section 3.21 for further discussion.) Inquiry into the behavior of the normal force when we press the book against the ceiling becomes a valuable homework exercise at this point, extending and enriching the context.

Since the normal force is usually first encountered at horizontal surfaces, still other subtleties behind the concept go unnoticed and unarticulated. Many students, in fact, interpret the word “normal” in its sense of “usual” or “ordinary” rather than its geometrical sense of “perpendicular.” The full meaning of the term does not become apparent until the confrontation with inclined surfaces, and, by this time, teachers frequently lose sight of the fact that the concept has not been convincingly explored, while many texts seem to take the attitude that it is too obvious to require discussion.

In my own experience, the physical situation shown in Fig. 3.19.1 is very helpful in raising and settling a good number of the issues involved. This apparatus is very widely used in showing composition and decomposition of forces (with actual numerical data being taken), and it is found in most preparation rooms. I have rarely seen it used, however, for explicitly generating the “normal force” concept by showing that the inclined plane exerts a force perpendicular to itself in the absence of friction.

After one balances the cart in the direction parallel to the plank, one proceeds to “replace,” by loading the second string, the force exerted on the cart by the plank. Many students do not notice the direction of this string unless the direction is explicitly called to their attention. They must also be

led to state the relationship between the force now being exerted by the string and the force previously being exerted by the plank.

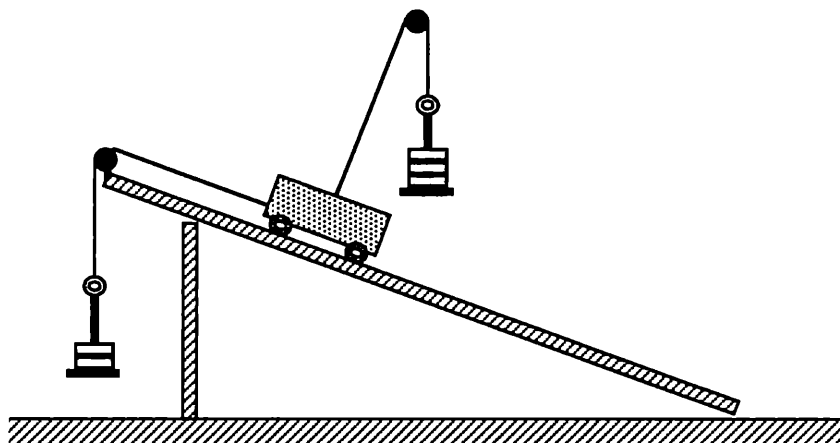


Figure 3.19.1 Demonstrating that the so-called “normal” force at an interface is indeed normal to the interface.

A powerful impression is then made by shifting the cart up the plank (and then down the plank) so that the string is visibly inclined from the perpendicular, and watching the cart oscillate while returning to the position at which the string is again normal to the plank. A gasp is frequently heard when this demonstration is performed, clearly indicating that the observed effect was unexpected. In most cases, in order to get all the relevant ideas fully registered, it is necessary to continue the discussion as far as examining the components of force, and the accelerating effect on the cart, when the cart is displaced from the equilibrium position.

This demonstration is valuable for at least two other reasons: (1) It constitutes an analogy for the concept that electrical field strength must be normal to the surface of a conductor; otherwise charge would be displaced along the surface until the field lines did acquire the normal orientation. (2) It helps students acquire a better understanding of *orthogonal components* of forces, a matter that will be discussed in more detail in Section 4.3.

3.20 OBJECTS ARE NOT “THROWN BACKWARDS” WHEN ACCELERATED

Consider the following situations: (1) a ball is placed on a cart, and the cart is accelerated from rest; (2) a pendulum bob hangs from the roof of an accelerated car; (3) a person is sitting in a car that begins to accelerate.

If asked about any one of these cases, a great many students contend that the person, the bob, the ball are “thrown backwards” when the vehicle

accelerates and, if asked to draw force diagrams, they show a force acting in that direction. The source of the difficulty is, of course, a very natural and common sense one: There is a strong inclination to put oneself into the accelerating frame of reference. These situations are far from trivial, and it is a mistake to consign them entirely to homework. At least one such situation should be discussed, with demonstration, in class.

No amount of previous discussion and definition of inertial frames of reference makes much impression on the majority of students until they encounter a noninertial frame and start confronting contradictions. In order to understand what an inertial frame is, one must begin to understand what it is *not*, and situations such as those proposed above are a first opportunity to make this point in rectilinear dynamics.

In the case of the ball on the cart (which can be assigned as a home experiment), most students are surprised to see that, although the ball rolls backward with respect to the cart, it moves forward with respect to the ground.

In the case of the pendulum bob, an excellent and very simple demonstration can be made by accelerating, in one's own hand, the top end of the string on which the bob hangs. Students can see the suspension point move forward while the bob retains its position relative to the floor. They can begin to discern that the bob is not thrown backwards relative to the floor and that acceleration of the bob begins only when the force exerted by the string acquires a nonzero horizontal component. (At this stage, many students are still very shaky about components of force and their accelerating effects, and this demonstration is particularly valuable because it invokes the concept of components in addition to frames of reference.)

Having examined cases (1) and (2) from the point of view of a bystander, students can now take up case(3) in which they are participants, as individuals in the accelerating car. They should be led to recognize explicitly that they are not thrown backwards but feel the force exerted on them by the back of the seat as the seat is accelerated—just as the pendulum bob experienced neither horizontal force nor acceleration until the inclined string began to pull it horizontally.

Not only does qualitative examination of these cases give students the opportunity for some valuable phenomenological thinking, helping them absorb the frame of reference concepts, but it also paves the way to better eventual comprehension of centripetal force and circular motion. The fact that time elapses between the two encounters is of vital importance, being conducive to learning.

If the teacher desires to do so, and if it is appropriate for the level of the students, the concept of “fictitious forces” can be introduced at this juncture. One of the very best presentations of these ideas is still to be found in the old PSSC film “Frames of Reference,” however dated it may seem to be.

3.21 FRICTION

Friction is a “passive” force in the sense defined in part 2 of Section 3.12; it adjusts itself in response to active effects. In fluids, the frictional resistance varies with the velocity of the moving object. At an interface between solids prior to slipping, the frictional force starts at zero and, as the force tending to produce slipping increases, the frictional force increases until the interface “breaks” and slipping begins. I use the word “breaks” not in a literal sense but to emphasize the analogy between this situation and that in which bodies literally do break under loading as the normal force increases to a critical value—as in the case of piling weights on a table until it breaks. This is an analogy that students do not perceive unless it is made explicit; yet, when it is established, they acquire a better understanding of the nature of the effect.

That such understanding is initially lacking in many students becomes evident if one observes some of the things they do in attacking end-of-chapter problems. They tend to use the formula $f = \mu N$ for any and every frictional force whether slipping is about to occur or not. In other words, they do not explicitly realize that the frictional force might have any value between zero and the maximum referred to in the formula. It is not that the textbook has failed to present the formula properly; this is competently handled in most books. The trouble is that the student has not been led to confront cases in which the value of the frictional force lies between zero and the maximum and thus fixes only on the formula. As in many other instances (e.g., the kinematic equations for uniformly accelerated motion), the student must be helped to see when an equation *does* apply and when it does *not* by dealing explicitly with cases in which it is inapplicable.

Another situation in which a force of static friction builds up from zero to its maximum value is that in which a frictional force acts to *accelerate* a body, as in the case of a block resting on an accelerating cart. The frictional force exerted on the block by the floor of the cart increases as the acceleration of the cart increases. Since there are no other horizontal forces acting on the block, this situation is fundamentally very different from the one in which a block is acted on by an external horizontal force while resting on a stationary platform, and many students have serious difficulty drawing a correct force diagram. Such situations are frequently encountered in end-of-chapter problems, but many students never acquire an understanding of the physics; they either never solve the problems correctly or they memorize procedures in which they plant μN 's around without understanding what they are doing. It is most effective to develop and contrast the two situations (block on the floor and block on the accelerating cart) when the concept of the static coefficient is first being developed. Enlarging the context for the same concept is conducive to learning and understanding.

As pointed out in Section 3.19, many students pick up the misconception that a normal force N is always equal to mg because they first encounter the

normal force in cases such as that of objects resting on horizontal floor or table surfaces with no vertical forces acting other than the weight mg . This subsequently leads to their treating every frictional force as being equal to μmg regardless of what the normal force actually is.

An effective way of displacing this misconception is to examine the situation of the book pressed against the wall, where the normal force has no connection whatsoever with the weight of the book (see Section 3.19.) The problem should be posed as one requiring investigation of changes (not just as a single calculation with one set of given numbers): (1) Draw force diagrams of both the book and the wall. (2) Suppose the horizontal force we exert on the book is very *large*: What are the magnitudes of the frictional force and of the normal force? How is the frictional force related to the normal force under these circumstances? How is it related to the weight of the book? (3) Suppose we start decreasing the horizontal force we are exerting: What happens to both the frictional force and the normal force as the decrease proceeds? Under what circumstances does the book begin to slip downward along the wall? How is the frictional force related to the normal force once sliding begins? How is it related to the weight of the book?

Textbooks and teachers frequently tell students that “frictional forces always oppose motion” without examining this phraseology critically. Students interpret the word “motion” in this context as referring to motion of the *body* on which the frictional force acts, and, in this sense, the statement is not always true. It is true that frictional forces at solid interfaces always oppose *slipping of the surfaces*, but in many instances of everyday experience the frictional force is the one that accelerates the body under consideration: The frictional force exerted on our shoe by the ground accelerates us when we walk; the frictional force exerted by the road on the tires accelerates the car; the frictional force exerted on the block by the floor of the accelerating cart (in the illustration discussed above) accelerates the block.

Many students initially have quite a bit of trouble in visualizing the direction of frictional force on each of two objects at an interface. When this is the case, I find the following approach helpful: I suggest that they put their two hands together, palm to palm, and imagine one hand to be one of the two objects and the other hand the other. Then I suggest that, concentrating on each hand in turn, they slide one hand over the other in the direction in which the objects would tend to slide, feel the force exerted on the hand, and put that force on the corresponding object in the force diagram. The extent to which students find this device helpful is evident when one sees how many are rubbing their hands over each other during tests.

3.22 TWO WIDELY USED DEMONSTRATIONS OF “INERTIA”

Two excellent demonstrations are widely used to demonstrate what is frequently (much too casually) described, as “inertia”:

- 1 The tablecloth is yanked out from under a set of dishes, leaving the dishes on the table.
- 2 A massive block is suspended by a string from a rigid support, and an identical string hangs from the bottom of the block. When the lower string is pulled slowly downward, the upper string breaks; when the lower string is jerked downward, the lower string breaks.

There is much more involved here than just “inertia.” Both of these situations are rich in physical phenomena, and students should be led to think about them in some detail in order to understand what is involved. Probably the best way to induce this thinking is to perform the demonstrations and ask enough leading questions (assigned as homework) to make it possible for the majority of students to fill in the gaps without getting bogged down.

In demonstration 1, if the dishes were glued to the tablecloth, they would be yanked off the table. The demonstration depends on the fact that the coefficient of friction is sufficiently low to allow the interface to “break” (in the sense defined in Section 3.21 above) at a value of maximum frictional force sufficiently small to impart sufficiently small acceleration to the dishes. Even with a relatively small frictional force, however, the dishes would still be yanked off the table if the tablecloth were very long, extending down the table well beyond the dishes. In other words, there is a time element involved, and the demonstration works because the time during which acceleration is imparted is short enough to make the displacement negligible.

The inertia of the dishes is indeed an important factor, but so are the others. Viewers of this demonstration are rarely given the opportunity to think it through and understand it fully. Part of the understanding depends on awareness of what *might* happen, of what is *not* the case—in addition to an awareness of what is the case and what *does* happen.

In demonstration 2, the crucial physical effect is the *stretching* of the strings to their breaking point. The stretching eludes direct sense perception and therefore has to be discerned in the imagination. Few students perform this act of imagination spontaneously, but it is not difficult to guide them into it. The key is again the element of time (as in demonstration 1, but in a somewhat different fashion): When the lower string is jerked, the low acceleration of the block allows the lower string to be stretched to breaking point before displacement of the block produces comparable stretching of the upper string; when the lower string is pulled slowly, both strings stretch without appreciable time delay, and the upper string is stretched to breaking point first because of

the higher loading. [At a still higher level of sophistication, students could be encouraged to visualize the elastic waves that must propagate up and down through the components of the system preceding the displacements leading to breaking. See part 1 of Sect. 3.12 above and the accompanying footnote.]

Without visualization of the stretching of the strings, students acquire no understanding of the demonstration; they simply memorize, and repeat, that it had something to do with "inertia."

3.23 DIFFERENT KINDS OF "EQUALITIES"

A hidden source of confusion for many students, one rarely recognized and eliminated in course work, is the fact that the "equals" sign ($=$) means very different things in different contexts. Following are some examples:

Statements such as

$$\rho = \frac{M}{V} \quad \text{and} \quad \bar{v} = \frac{\Delta s}{\Delta t}$$

are actually definitions (or identities) rather than ordinary functional equalities, and one should use the three-line symbol (\equiv) for "defined as" or "identical with" rather than the ordinary equals sign. (Some texts are now doing this, but the reason must still be discussed and emphasized to the students.)

The kinematic equations, however, are statements of functional equality (subject to the restriction to rectilinear motion and uniform acceleration) *derived* from the *definitions* of s , t , v , and a ; they are like the equations the students have become familiar with in elementary algebra. The ordinary equals sign ($=$) is appropriate.

The equals sign in $\vec{F}_{\text{net}} = m\vec{a}$ is not just an ordinary functional equality. It conceals the combination of arbitrary definition and laws of nature lying behind either the Machian or Newtonian approach to the second law (see Sects. 3.9 to 3.6). One side cannot replace the other in a force diagram.

The statements $f_{\text{max}} = \mu N$ and $F = kx$ (Hooke's law) are of only limited validity and applicability. Hence, the equals sign applies only under certain restrictions, the text of which must accompany the symbols.

The impulse-momentum and work-kinetic energy theorems, when derived from $F_{\text{net}} = ma$, reveal the remarkable and unanticipated equality of numbers calculated in entirely different ways.⁸ One way involves the necessity of

⁸From a still more advanced point of view these theorems begin to exemplify the difference between a line integral (or inexact differential) on the one hand and a state function (or exact differential) on the other, and pave the way for more sophisticated utilization of these ideas, for example, in thermodynamics. It is not appropriate to belabor this mathematical aspect with students who are not ready for it and who are not going on to more advanced physics courses. It is something well worth returning to, however, from the perspective of a more advanced course and, if the foundation has been started in an elementary way in the introductory course, understanding of the more advanced and subtle ideas is greatly enhanced.

knowing the history of variation of F_{net} as a function of either clock reading or position over the entire interval in question and finding the area under the appropriate graph. The other way involves only the initial and final velocity state of the body over the given time or position interval. The equations are therefore not merely functional relations; they have physical content and meaning not articulated in the equals sign alone.

A common, but nevertheless confusing, use of the equals sign is in statements such as $2.00 \text{ m} = 6.56 \text{ ft}$ or $1 \text{ kg} = 1000 \text{ g}$ or, even worse, $1 \text{ kg} = 2.20 \text{ lb}$.

These, of course, are not equalities at all but describe various kinds of *equivalence*, requiring an accompanying text of interpretation. Students recognize that the numbers shown in the statements are clearly not equal, but many are afraid to ask about the contradiction, sensing that the question is likely to be regarded as “stupid.” Furthermore, it is difficult and confusing for many students to see that the statement $6P = S$, where P and S are *variables*, is profoundly different from the statement $2.00 \text{ m} = 6.56 \text{ ft}$, where m and ft are *not* variables. Once they have mastered the meaning of something like $6P = S$, they are easily misled into thinking that the other relation implies that 2.00 times the number of meters is equal to 6.56 times the number of feet. (See Section 1.15 for additional detail.) Some equivalence symbol, distinctly different from the equals sign, would be much more appropriate in this context.

Since the same symbol ($=$) is used throughout these different contexts, students are not impelled to discern the profound differences in meaning unless the differences are discussed explicitly. One of the consequences of omission of such discussion is, for example, the notion that “ ma is a force since $F = ma$.” The result is the appearance of a force, labeled ma on the force diagram of an accelerated car or on the force diagram of a person seated in the car. Another consequence is the confusion attending translation of verbal statements into symbols that was discussed in Section 1.15.

These differences in the meanings of equals signs are, of course, not confined to physics; they permeate arithmetic and mathematics as well, and, if made clear in both areas, help develop student understanding that much more rapidly.

An analogous problem that is rarely noticed by teachers has to do with the concept of “zero.” The term is used, with different meaning, in several different contexts. To students it frequently means little more than “absence of anything.” In another context it might be a “starting level.” On other occasions, it is a number but is not recognized as such. In still another context, it is a position along the number line. Students should be made explicitly aware of these different levels of meaning and interpretation.

3.24 SOLVING PROBLEMS

The importance of helping students acquire the habit of a systematic approach to solving end-of-chapter problems has been mentioned in Sect. 2.13 and, if such habit has been cultivated in kinematics, it provides a strong foundation for progress in the more subtle area of dynamics. Many textbooks give specific examples of a systematic approach, but many students resist using it unless they are required to do so as part of graded performance.

Guided drill sequences can be very helpful to students who find difficulty implementing the scheme outlined in their text. Such drill can be provided in cooperative learning sessions [c.f., Heller et al, (1995)] or by computer-based dialogues. Students should be led to draw force diagrams first, set up a coordinate system, apply the second law, and carry out solutions. Not only does the systematic approach facilitate solving the problem; it also helps break down the very high resistance, among many students, to putting pencil to paper, and thus beginning to analyze the problem, before the entire solution or “answer” has been “seen” without analysis.

One example of a systematic aspect of problem solving in elementary dynamics that is omitted by many textbooks and programs is the explicit writing out of the equation for continuity of acceleration when a problem deals with two or more interconnected bodies. In the case of the Atwood machine, for example, the symbol a is casually written down for both bodies without explicit recognition that the equality of the accelerations of the two bodies is one of the essential mathematical conditions describing the physical system (i. e., the nonstretching of the string).

This omission then causes students great difficulty if they confront more complicated problems (e.g., an Atwood machine suspended on one side of another Atwood machine) in which an equation relating the various accelerations is an essential part of the solution. Omission of the acceleration equation also makes students lose sight of the fact that, in the Atwood machine, for example, the two accelerations are *not* the same if the string is stretching. They are also kept unaware of the fact that the derived results do not apply to the short but finite time interval during which the waves bounce back and forth along the string after the system is released. Greater care in making the acceleration condition explicit prepares students going on to more advanced physics for the very careful equation writing they must do in more sophisticated theoretical analyses.

Many textbooks fail to include certain valuable aspects of problem solving that can quite readily be added by interested teachers:

- 1 Problems should occasionally include information irrelevant to the solution. Part of understanding a physical situation, and of solving a problem concerning it, is being able to discriminate relevance and irrelevance. Such discrimination requires practice (it does *not* develop

spontaneously), but students are very rarely given such practice. When suddenly confronted with irrelevant information, many students force it into their problem solution to make sure they have “used all the data.” Then, when made to realize the irrelevance, they complain about “difficulty” and “unfairness” with no realization that, in real problems, they will have to winnow relevance and irrelevance on their own. The point and purpose of injecting irrelevant information should be discussed explicitly; students are perfectly willing to accept such practice once they understand what it has to do with their own intellectual development.

- 2 Problems, where appropriate, should require development of a *complete* algebraic solution for the unknown quantity before substitution of any of the given numerical values. Many students initially do not understand what is meant by “complete algebraic solution,” even if this is demonstrated in text or lecture. When called upon to set up such a solution themselves in a new problem, they believe that any algebraic relation they write down as a start satisfies the requirement. There is tremendous resistance to continuing through several steps of combining algebraic equations and solving for unknowns prior to making numerical substitutions. This resistance can be reduced only by fostering practice and, at least in some instances, forbidding numerical substitutions at intermediate points in a solution.
- 3 Another skill that requires practice, and which is inadequately cultivated in textbooks, is that of interpreting both numerical and algebraic problem solutions in words. If this is to be done, it must be elicited by the teacher in both tests and homework since it is not fostered elsewhere. This is a rather sophisticated intellectual mode, and very few students develop it spontaneously, but they show gratifying progress with practice—gratifying to themselves as well as to the teacher.

In the initial stages and without the practice, a correct numerical or algebraic result does not necessarily indicate understanding on the part of the student. When one calls for, and examines, the interpretation, one frequently finds very serious errors, misapprehensions, gaps in understanding, and failure to perceive some of the content and implications of the solution.

Interpretation of numerical results should include an assessment of whether or not the order of magnitude makes sense, with supporting argument. Interpretation of algebraic solutions should include examination of extreme, limiting, or asymptotic cases, especially when these limits show whether or not the result makes sense and thus constitutes a check on the correctness of the solution.

Problem solving as an aspect of student cognitive development and performance is a subject of active research in many fields and on various fronts.

There is a large and rapidly growing literature. No attempt will be made to discuss this area of investigation in detail in this book, since the concentration here is on the formation of underlying concepts that *precede* application in problem solving rather than on problem solving itself. In fact, my own experience indicates that, given careful building up of the underlying concepts, cultivating sensitivity to definition, and enhancing verbal expression, make many of the difficulties now encountered in problem solving go away.

Readers interested in gaining entry to the problem-solving literature relevant to physics teaching will find good starting points in Lapp (1940), Larkin (1981) and (1983), and Reif et al, (1981), (1982), and (1984). An especially fine article of broad generality, with many specific examples, is available in Reif (1995).

3.25 SAMPLE HOMEWORK AND TEST QUESTIONS

1 Suppose you are sitting on a chair that stands on the ground. Draw well-separated force diagrams of your body, the chair, and the whole earth. Describe each force in words [describing in words means indicating the nature of the force (gravitational, contact frictional) and stating what object exerts that force on what.] Show the relative sizes of the forces by using a longer arrow for a larger force and equal-length arrows for forces equal in magnitude. Identify the third law pairs.

2 Suppose you are standing on the ground in a shed and pulling vertically downward on a string that is attached to the bottom of a block that hangs from the ceiling on a rope. Draw well-separated force diagrams for your body, the string, the block, the rope, the shed, and the whole earth. Do all the various things called for in Question 1. Now repeat the exercise for the case in which you pull the string at about 45° from the vertical.

3 Suppose you are in the act of jumping vertically upward. Your legs are flexed and pushing on the floor so that your body is being accelerated upward.

- (a) Draw well-separated force diagrams of your body and of the earth. Show the relative magnitude of various forces; describe each force in words; identify the third law pairs.
- (b) Draw the force diagrams for the situation that obtains just after your body leaves contact with the floor on your way up in the jump and do the other things called for in (a).
- (c) Repeat for the situation at the top of the jump, for some point on the way down, and for the situation just after you hit the ground, and your bent legs are slowing you down.

4 Suppose you throw a ball vertically upward and catch it when it returns. Draw force diagrams for your body, the ball, and the earth (a) while you are accelerating the ball upward; (b) while the ball is rising after having left your hand; (c) while the

ball is falling; (d) while you are in the act of catching it. Describe forces in words; show relative magnitudes; identify third law pairs

5 Suppose you start walking or running. In either case you start with an initial velocity of zero and accelerate to a nonzero velocity in the horizontal direction. This means that there must have been an unbalanced horizontal force acting on you during the acceleration.

- (a) Proceed to analyze and sense what is going on by actually performing the actions. As you do so, sense the direction of the horizontal force acting on *you*; then pretend you are the ground and visualize the direction of force you would feel. Now draw well-separated force diagrams of both you and the earth. Describe each force in words and identify third law pairs.
- (b) How do the diagrams you have drawn in (a) differ from the ones when you are standing still? Label, on the appropriate diagram, the force that imparts acceleration to your body. What is the role of friction in this system? Could you walk or run in the absence of a frictional force between the soles of your shoes and the ground? Why or why not? Discuss, in terms of the force diagrams, what happens when you try to walk on an icy surface.
- (c) Discuss the following statement (i. e., is it correct and accurate or is it incorrect? Explain your answer and, if you believe the statement to be incorrect, alter it in such a way as to make it correct.): “When we walk or run in the forward direction, we push on the ground with a horizontal frictional force directed toward the rear. The ground, in turn, pushes on us with a horizontal frictional force in the forward direction. We are accelerated by the force exerted by the *ground*, not by the force that *we* exert.”

6 Following the same sequence as in Question 5, analyze the forces that act when a car accelerates along a road. (If you have available a spring-wound or electrically driven toy car, place it on your hand and sense the direction of the force exerted on you by the wheels as the car accelerates from rest along your hand.) Give a careful verbal description of the force that accelerates the car: How does this force originate and what object exerts it on what? Is it correct to say that “the car is accelerated by the force exerted on it by the engine”? Explain your answer.

7 A railroad car is in *uniform* rectilinear motion along its track. Observer A performs experiments inside the car while Observer B, outside the car at a fixed location along the track, watches these experiments as A comes by. In each of the following experiments with various objects, describe what each observer will see the object doing relative to his *own* frame of reference (i. e., how will A see the object behaving relative to the inside of the car, and how will B see it behaving relative to the ground?). As part of your description be sure to sketch a diagram of what each observer sees to be the trajectory of the moving object.

- (a) Observer A puts a ball down on the perfectly level floor of the car. (Will the ball stay wherever he places it?)
- (b) Observer A makes the ball roll on the floor in a straight line at uniform velocity relative to himself: (1) toward the front of the car; (2) toward the back of the car; (3) directly across the car.

- (c) Observer A lets the ball fall out of his hand from a point several feet above the floor of the car.
 - (d) Observer A throws the ball vertically upward and catches it as it comes down.
 - (e) Observer A suspends a pendulum bob on a string from the roof of the car. How will the pendulum appear to hang?
 - (f) Observer A throws the ball with an initial horizontal velocity directed toward the front of the car.
 - (g) Observer A throws the ball with an initial horizontal velocity directed toward the rear of the car. (As part of this problem, consider the important special case in which the *magnitude* of this horizontal velocity is equal to the magnitude of the horizontal velocity of the car relative to the ground.)
 - (h) Observer A has an aquarium tank of water resting on the floor of the car. (How does the surface of the water behave—is it level or sloping?)
- 8 Consider once more all of the situations in question 7, but the car now has a uniform *acceleration* in the forward direction. Describe, with appropriate diagrams, what each observer sees relative to his frame of reference. What would be the difference between the personal sensations experienced by A and B?
- 9 Suppose you are sitting on a platform that is subject to very small frictional forces—for example, a large frictionless puck on a huge air table or a boat floating on water. You are initially at rest, and you have no paddles and no way of pushing yourself along the surface, but you wish to accelerate yourself and your platform to a finite velocity in some particular direction. You have available a basket of balls or stones
- (a) What might you do to accelerate yourself in the direction you wish? What might you do if you wish to slow down or stop? Analyze this situation carefully, drawing well-separated force diagrams for a ball, yourself, and the platform. Explain, in terms of Newton's laws, what is happening to make your transportation possible.
 - (b) In the light of the thinking you have done in part (a), analyze, with appropriate force diagrams, what must be happening (1) when you paddle a boat through the water; (2) when a propeller accelerates a boat or an airplane. Why is the action of the propeller still necessary once the boat or plane has been accelerated to desired velocity? Why is a propeller useless on a space ship?
 - (c) In the light of the thinking you have done in part (b), describe what must be happening in rocket propulsion of a space ship. What happens to the fuel that is burned in a rocket engine? Draw well-separated force diagrams for the ship and for a parcel of gas (resulting from the burning of the fuel) being ejected from the ship. What must be done to slow down the space ship? What must be meant by the technical term “retro-rockets”?
- 10 We have given the name “weight” to the gravitational force exerted by the earth on all objects, including our own bodies. Note, however, that we do not have any direct feeling or sensation of the force of gravity—the pull of the earth on us. As

we stand on the ground or sit on a chair, what we feel is not the pull of the earth but rather the upward force exerted on us by the ground or by the chair. When we jump from a height and are falling freely, our principal sensation (apart from the rushing air) is the *absence* of the upward force to which we are so accustomed, not a downward pull.

- (a) Verify the preceding statements by consciously examining your own sensations in these various circumstances: Standing, sitting, jumping vertically upward, or jumping off a chair. (The interval during which you are out of contact with a support when you jump is very short, but you can still check on the fact that you do not feel anything pulling on you.)
- (b) Now suppose you are standing on a bathroom scale in an elevator that is standing still. Draw force diagrams of yourself, of the scale, and of the floor of the elevator where it interacts with the scale. How does the upward force exerted by the scale platform *on you* compare in magnitude with your weight? How do you interpret the reading on the scale? (Explain your answers in your own words.)
- (c) Suppose that, instead of standing still, the elevator is moving up or down with *uniform* velocity. Draw the same diagrams called for in (b), and answer the same questions. When the elevator is moving up or down at uniform velocity, how does the total upward force exerted by the cables on the elevator compare with the total weight of the elevator and its contents? (Explain your answer.)
- (d) Suppose that you are still standing on the scale in the elevator, and the elevator is *accelerating upward*. Stop and recall the sensations you have felt during the short interval when an elevator accelerates upward before attaining a uniform velocity. Draw the same force diagrams called for in (b). How must the magnitude of the upward force exerted by the scale platform on you compare with your own weight? How must the reading on the scale compare with the reading you make when the elevator is standing still or moving up or down at uniform velocity? Is your conclusion about the size of the upward force exerted on you by the scale consistent with the sensation you experience during the interval of upward acceleration? How must the total upward pull of the cables on the elevator compare with the weight of the elevator and its contents? What happens to the magnitude of the reading on the scale and the total upward pull of the cables as the acceleration is made larger and larger? (Explain your answers.)
- (e) Suppose you are standing on the scale in the elevator and the elevator is accelerating *downward* with an acceleration smaller than 9.80 m/(s)(s) . Go through exactly the same sequence of diagrams and questions as you did in part (d) including visualizing what happens as the downward acceleration is made larger and larger.
- (f) Suppose the upward pull of the cables on the downward accelerating elevator is made smaller and smaller until the elevator is falling freely. Under these circumstances what forces are acting on *you*? What is the magnitude of the upward force exerted On you by the scale platform? What is the numerical reading on the scale? Has your weight (the gravitational force exerted on you by the earth) changed?

- (g) The elevator is completely enclosed, and you cannot see out; your frame of reference is the interior of the elevator. Describe what you would observe happening in the freely falling elevator as you performed various simple experiments such as: (1) Holding a ball in your outstretched hand and letting it go; (2) pushing the ball off gently in various directions (up, down, sideways, etc.); (3) suspending an object on a spring balance; (4) attaching a pendulum bob on a string to the ceiling of the elevator and trying to swing the pendulum. What would your own sensations be like in the freely falling elevator? Could you “stand” on the floor as you do in ordinary circumstances? What would happen if you pushed yourself away from the floor or a wall? Invent some additional questions and thought experiments of your own.
- (h) Suppose you are inside a capsule or elevator that is projected upward by a powerful catapult. (1) Describe the sensations you would have and the reading you would observe on the platform scale during the interval in which the catapult is projecting you upward. (2) Describe the conditions you would observe in the capsule as it is rising vertically after having left the catapult. How do these conditions compare with those you described in the freely falling elevator?
- (i) Is it, in principle, possible to accelerate an object downward with an acceleration larger than that of free fall, that is, larger than 9.80 m/(s)(s) ? What would have to be done to impart such an acceleration to a stone or to an elevator? What would conditions be like in an elevator accelerating downward with an acceleration larger than g ? Repeat some of the thought experiments suggested in part (g) and describe what you would observe. Where would you be likely to find yourself standing during the interval in which such an acceleration is maintained?
- 11** Suppose you are sitting in a car that is speeding up. Draw well-separated force diagrams of the following objects: your own body; the seat in which you are sitting (apart from the car); the car (apart from the seat); the road surface where the tires and road interact. Assume the car has rear-wheel drive. Describe each force in words; show large forces with longer arrows; identify the third law pairs. Explain carefully in your own words how the force imparting acceleration to the car originates.
- Suppose a pendulum bob hangs from the ceiling of the car. Draw well-separated force diagrams for: the bob; the string on which it hangs; the car ceiling where it interacts with the string. Describe each force in words; identify the third law pairs.
- 12** Suppose you are sitting in a moving car that is slowing down under the influence of four-wheel brakes. Do for this case all the things called for in question 11. Explain carefully in your own words how the force slowing down the car originates.
- NOTE: The preceding questions are designed to start with situations in which the student is called upon to analyze his or her own bodily sensations in terms of the dynamical concepts being developed. Insights so formed are then extended to situations that can only be visualized in the abstract or to unfamiliar situations that are “transformations” of the familiar ones.
- 13** Consider a pendulum bob on a string attached to the ceiling. The bob is pulled off to the left and let go. Sketch the trajectory (the path) that would be followed by

the bob in each of the following instances: (1) The string is cut at the instant the bob is about halfway down to the lowest point in its swing; (2) the string is cut at the instant the bob has reached the lowest point in the swing; (3) the string is cut at the instant the bob is about halfway up to the highest point it would reach on the right; (4) the string is cut at the instant the bob is just at the highest point of its swing on the right.

14 In connection with his development of the concept of universal gravitation and the role of mass in the inverse square law, Newton did a highly significant experiment: He constructed a hollow pendulum bob into which he could put objects made of different materials, such as wood, iron, gold, copper, salt, and cloth. One can measure the period of a pendulum very precisely by counting swings over a sufficiently long time. Newton found that the period of the pendulum did not change observably with very different materials inside the bob. Why did Newton feel it important to do this experiment? What ideas does the null result support?

NOTE: Although many textbooks include qualitative questions intended to evoke some of the kinds of thinking described in this chapter, most fail to achieve their desired end for several reasons: The questions are fragmented and isolated in such a way that they do not help the student form a synthesis of the ideas under consideration; the questions are too sophisticated and too abrupt and do not provide sufficient Socratic guidance to the student who is having difficulty; there is insufficient spiralling back and repetition in subsequent, richer context. Apart from weaknesses in design, the questions also have very little effect because they are rarely, if ever, invoked in testing.

In addition to questions of the type illustrated above, most of the questions used in research on student learning, and described in the body of this chapter, are useful in helping students make the errors and encounter the contradictions that lead them to refine their conceptions and master the insights we wish to instill. These questions supplement and make more effective the conventional end-of-chapter problems provided in textbooks. The questions have very little effect, however, if they are not assigned, discussed, and included in testing and grading.

Chapter 4

Motion in Two Dimensions

4.1 VECTORS AND VECTOR ARITHMETIC

Most presentations of the concept of a vector start with the representation of displacements in two dimensions and develop the process of addition of such quantities in an intuitive way. This is, without question, the most reasonable and effective starting point, and students have relatively little difficulty with the ideas in the early stages. Trouble begins to set in, however, when abrupt jumps are made to other operations and to other vector quantities.

To generate the process of subtraction, for example, some textbooks simply assert that, to obtain the negative of a vector, one reverses the direction of the original arrow (or one simply multiplies by the scalar factor -1). Although the reason is quite obvious to us, it turns out to be far from obvious to many students; they hesitate to ask for a reason, and they memorize the assertion without understanding. By the time one wishes to find the vector change in velocity over a small angular displacement in circular motion, for example, the process has been forgotten, and the derivation of centripetal acceleration is not understood.

A more effective way of introducing the operation of subtraction is to adopt the systematic procedure of mathematics and ask what must be added to a given vector to obtain a zero vector. (In the technical terminology of mathematics, one is generating the “additive inverse.”) This serves to define subtraction by tying it to the original starting point, namely addition, and gives the student logical continuity rather than abrupt change and unsupported assertion.

Many texts go on to assert that, since they also have “magnitude and direction,” velocity, acceleration, and force are also vector quantities in the sense of obeying the same arithmetic as displacement. This, again, is far from obvious to the students; to them, adding velocity arrows appears very different from adding displacement arrows, and acceleration arrows are totally incomprehensible. The transition from one kind of quantity to another requires

some discussion if vectors and their arithmetic are going to be used. It is not, in fact, “easier” for the learner if the logical questions are concealed or ignored.

To do this honestly requires at least some discussion of multiplying displacement vectors by constants larger and smaller than unity, of what happens when one divides the displacement by the time interval in which the displacement occurred, of units and dimensions of such quantities, of acceleration as a vector, and of the connection between force and acceleration. (It should be noted that the path is much smoother if attention was given to the algebraic signs of acceleration in rectilinear motion, as advocated in Section 2.8, and to careful operational definition of force as suggested in Sects. 3.4 to 3.6.)

Another property of vectors, frequently taken for granted in instruction without being made explicit, is that of “movability.” Many students tenaciously hold an initial view that vectors are “attached to points.” One can see how this notion gets planted: Displacements, the first vectors encountered, begin at a fixed position, and a sequence of displacements proceeds from point to point with the arrows head to tail, each tail rooted at the initial fixed position; velocity vectors appear to be attached to particles; concentrated forces act on objects at a point. When it comes to recognizing that forces can be considered as acting anywhere along their line of action or to transferring velocities from the diagram of circular motion in order to draw the vector diagram for change in velocity, one encounters resistance and disbelief. One is doing things that “don’t make sense,” are not understood, and are therefore avoided, or handled incorrectly, in the solving of problems on tests and homework. The fact that vectors are not attached to points requires explicit discussion if it is to be understood—and used in attacking problems.

Many students would benefit from more exercises and drill in graphical handling of vector arithmetic than are usually available in textbooks. Such drill, with immediate feedback and reinforcement, could well be supplied through computer-based materials exploiting currently available graphic capability. Unfortunately, at the time of writing, very few well-designed materials seem to be available commercially.

4.2 DEFINING A “VECTOR”

How rigorously one should pursue the definition of “vector” is a matter of judgment, and I myself do not believe that it should be pressed very hard in lower level courses or with students who are not going on to more advanced mathematics and physics. For the few front runners, however, and for students likely to continue as physics majors and engineers, this question should not be ignored or glossed over.

A first step is to show students that just saying “magnitude and direction” is not sufficient to define a vector. Finite angular displacements, for example, do have both magnitude and direction but are not vector quantities because they do not commute on addition. This fact can be readily demonstrated for

students by having them rotate a book through two successive 90° displacements about two different axes and showing that the book winds up in two entirely different final orientations if the order of the rotations is reversed. This can be directly contrasted with the fact that displacements, velocities, and the like do commute when being added (or subtracted) and that this property is an essential part of the definition. (In order to understand what something is, one must also understand what it is *not*.)

Introducing the students to finite angular displacements has a very significant payoff later on when one wishes to show that angular velocity can be represented as a vector quantity. (This is usually asserted without any attempt at justification, but such unsupported assertions encourage memorization without understanding.) One can examine infinitesimal angular rotations to show that they do commute in the limit of zero displacement, and one is on the way to making the angular velocity vector a rational and plausible concept instead of a mystery to be memorized.

Another aspect worth bringing out is the mathematical view of vectors as “ordered pairs” of numbers. This ties in especially well, in due course, with the notion of rectangular components.

Commutation in addition and subtraction does not, of course, complete the definition of a vector. The final part of the definition resides in behavior with respect to transformation under rotation of coordinate axes, and this behavior is crucial to the final distinction between Cartesian and pseudo vectors. I find that an understanding of these distinctions, and of the need of extension of the basic definition beyond the requirement of commutation in addition and subtraction, comes far more easily to students in more advanced courses if they have the advantage of having been gradually exposed, in introductory courses, to the simpler ideas outlined above, instead of suddenly encountering all of them *de novo* at the advanced level. Such preparation, in other words, operates to give students a very much better grasp of the nature of the vector cross-product.

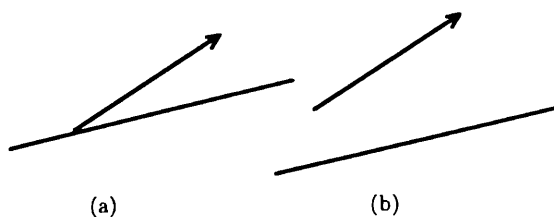
4.3 COMPONENTS OF VECTORS

The concept of orthogonal (or Cartesian) components of vectors seems so simple and transparent to teachers, and manipulations, when the Cartesian axes are given in a problem, are so easily memorized by students, that many significant student difficulties in this area go unnoticed. Interviews with students, however, reveal very significant gaps in understanding.

Consider the two diagrams shown in Fig. 4.3.1. If one draws diagram (a) and asks the student to “show graphically how large an effect the vector represented by the arrow (perhaps a force or a velocity) has along the direction indicated by the line,” many students find themselves at a loss and are unable to answer the question. If one draws diagram (b) and asks the same question, still more students are unable to answer. (In the latter case the difficulty

has been enhanced by the fact that the line does not pass through the tail of the arrow. See comments in Section 4.1 concerning this aspect.) Without an angle marked with a familiar symbol, with no Cartesian axes shown in familiar orientations, and with the word “component” not used, nothing triggers the student to bring out the memorized formulas with sines or cosines. In other words, students exhibiting this difficulty have not formed an understanding of the concept.

Figure 4.3.1 What is the magnitude of the “effect” of the vector in the direction indicated by the line?



There are quite a few ways of introducing concrete thinking and experience that help students assimilate the concept, and I do not see any particular one as superior to all others. It is necessary, however, to give the idea more attention than is accorded in most textbooks, and I have, over the years, narrowed down on the following mode in my own practice: Starting with an instance in which the arrow in Fig. 4.3.1(b) represents a displacement and the line represents a wall, we set up the corresponding situation in the classroom (or lecture room). The beam of an overhead projector is directed perpendicular to the wall so that the shadow of the displaced object (a student, a hand, a ball) is seen to be displaced along the wall while the object itself is displaced as indicated by the arrow. (We concentrate at this point only on the change of position and not on the motion.) We describe the displacement of the shadow as the “effect,” along the plane (or direction) of the wall, of the actual displacement of the object in the room. A similar observation is then made of the displacement of the shadow along the orthogonal wall.

We then investigate a few changes in order to enrich the context: What happens to the two “effects” when the original displacement is parallel to one wall and perpendicular to the other? Under what circumstances are the two “effects” equal in magnitude? What happens to each “effect” as the magnitude of the original displacement is kept fixed while the angle relative to one wall is increased or decreased? Must the axes along which we desire to measure “effects” be oriented only horizontally and vertically?

After such concrete experience, the students can quite easily be led to describe the corresponding pencil-and-paper procedure of dropping perpendiculars to the orthogonal axes and to recognize how the relevant sine and cosine expressions arise. Now that the idea has been fully established operationally, it is appropriate to introduce the technical name: “rectangular components of the displacement.”

In following stages, at a pace consistent with the readiness and sophistication of the students, one can examine (or visualize) the behavior of the two

shadows of a continuously moving object and extend the concept to include components of velocity and acceleration. Logical extension to the notion of components of force follows from whatever has been done in constructing the concept of “force” to begin with. For many students, however, it is an important part of concept building actually to pull a block (or some other object) along the floor or table with a string oriented at various different angles to the horizontal and to sense that their pull has two simultaneous but separate effects, one vertical and one horizontal, and that the magnitudes of these two effects vary with the angle of the string in essentially the same manner as did the magnitudes of the shadow displacements and velocities along the wall.

After one has built such an underlying structure, going back to a question such as that asked in connection with Fig. 4.3.1 (but with the picture rotated into an entirely different orientation) becomes a first test of whether the student has begun to absorb the concept.

As in the case of graphical addition and subtraction of vectors (Section 4.1), supplementary drill in interpreting components and in adding and subtracting vectors arithmetically by use of rectangular components is needed by many students. Computer-based materials could very effectively fill the gap usually left by textbooks and teachers for lack of space and time.

4.4 PROJECTILE MOTION

Examining Galileo’s own view of what he did in solving the problem of projectile motion and examining, at the same time, some of the logical and epistemological questions arising in the story provides an especially valuable opportunity to enhance the “scientific literacy” of the entire spectrum of students in introductory physics courses—from nonscience majors to future engineers and physicists. The degree of mathematical sophistication invoked can be very different for different groups, but the important ideas can still be brought out and discussed in an intellectually honest way with any group of students.

Galileo’s solution of the problem of projectile motion in the idealized limit of zero air resistance represents one of the very first deliberate uses of the concept of superposition in science, and, in the *Two New Sciences*, he describes his approach, and its conceptual importance, with utmost clarity:

In the preceding pages we have discussed the properties of uniform motion and of motion naturally accelerated. . . . I now propose to set forth those properties which belong to a body whose motion is compounded of two other motions, namely, one uniform and one naturally accelerated. . . . This is the kind of motion seen in a moving projectile; its origin I conceive to be as follows: Imagine any particle projected along a horizontal plane without friction. . . . This particle will move along this plane with a motion that is uniform and perpetual, provided the plane has no limits. But if

the plane is limited and elevated, then the moving particle, which we imagine to be a heavy one, will, on passing over the edge of the plane, acquire, in addition to its previous uniform and perpetual motion, a downward propensity due to its own weight; so that the resulting motion. . . . is compounded of one which is uniform and horizontal and of another which is vertical and naturally accelerated.

The key phrases in the preceding quotation are “resulting motion” and “compounded of.” With these deceptively simple terms, Galileo reveals how far he has come from the Scholastic point of view in which motion was seen only as a whole and never conceived as compounded. These phrases also articulate his inductive guess that the horizontal and vertical motions of a projectile do not influence each other, that is, that they behave as though each alone were present and that the net effect is a simple combination of the two independent motions calculated separately. This is a hypothesis about physics and is not just a matter of definition; verification is required—just as with the hypothesis that free fall is uniformly accelerated.

To justify the superposition, we must know the answers to two separate questions: (1) Does imparting a horizontal velocity to a particle in any way alter the vertical acceleration and velocities it normally acquires in free fall along a straight line? (2) Conversely, does the presence of a vertical acceleration and vertical velocity alter the horizontal velocity a particle might initially have? It must be emphasized that these are indeed two separate questions and that, if one is true, the converse does not automatically follow.

Galileo, of course, could not test these conditions experimentally with any great degree of precision (although he does claim that an object dropped from the top of the mast of a moving ship hits the deck at the base of the mast). He had to depend principally on internal consistency and overall agreement of the derived results with experiment. In this modern age of electronic devices and high-speed photography, it is possible to make direct tests that can be shown to the students. The widely reproduced stroboscopic photograph (all editions of *PSSC Physics* and many other textbooks) showing that two balls released simultaneously, one dropped vertically and the other having an initial horizontal velocity, occupy the same vertical levels at successive intervals of time, gives an affirmative answer to question (1). Films showing that an object dropped from a mast or vertical standard moving at uniform velocity falls directly along the mast and lands at its base give the affirmative answer to question (2). (The logical necessity of examining both of these questions is, unfortunately, glossed over in many presentations.)

To help register an understanding of the physics of projectile motion, it is important to have students draw *separate* diagrams of (1) the force acting on the projectile at various different points of the trajectory (e. g., a point on the way up, the top of the flight, a point on the way down); (2) the acceleration at

the same points; (3) the horizontal and vertical velocity components at each point; (4) the total vector velocity at each point. A similar array of diagrams should be drawn for the frictionless puck while it is moving along the table and for its trajectory after it flies off the edge of the table. Illustrations in textbooks sometimes have arrows for different quantities on the same diagram. This is invariably a source of serious confusion.

Still another valuable exercise is to sketch the trajectory that would be followed by the pendulum bob after its string is cut: (1) At some point in the downward swing; (2) at the lowest point of the swing; (3) at some point during the upward part of the swing; and (4) at the instant of the end of the swing. (The role and importance of such exercises was initially discussed in Section 3.16, and a statement of this problem is given in item 13 in Section 3.25.)

In the *Two New Sciences*, Galileo is doing much more than just presenting his mature insights into strength of materials and kinematics. He is also propagandizing the use of mathematics in the description and understanding of natural phenomena. After deriving the equation of the trajectory in projectile motion, he proceeds to show that the range must be a maximum at an angle of elevation of 45° .¹ His delight in this result is apparent in his having Sagredo say:

The force of rigid demonstrations such as occur only in mathematics fills me with wonder and delight. From accounts given by gunners, I was already aware of the fact that in the use of cannons and mortars the maximum range . . . is obtained when the angle of elevation is 45° . . . ; but to understand why this happens far outweighs the mere information obtained by the testimony of others or even by repeated experiment.

If the above quotation is used, it should be accompanied by a discussion of what scientists mean when they use the phrase “understand why. . .” in this and similar contexts. Students take this phrase far too literally unless it is explicitly qualified as referring to explanation *in terms of* an array of simpler, plausible, preferably well-established, ideas that underlie a very much wider range of phenomena but that themselves have no explanation and may have an endless regression of unanswered “why” questions behind them, as, for example, why all objects fall with the same acceleration regardless of their

¹For students who have gotten far enough with the calculus, this is a valuable opportunity to find the extremum by differentiation and to ascertain whether it is a maximum or a minimum; they have rarely had the experience of using this mathematical technique in a physical problem. For students who have not had calculus, it is perfectly possible to find the maximum in an intuitive, non-rigorous way. This is well worth doing both for the sake of appreciating Galileo’s argument concerning the role of mathematics and for the sake of connections with other physical situations in which similar maxima, associated with the product of the sine and cosine functions, occur.

weight or why the orthogonal components of projectile motion are independent of each other.

4.5 PHENOMENOLOGICAL THINKING AND REASONING

The purpose of this section is to illustrate certain kinds of thinking and reasoning that few students enter into spontaneously but that can be cultivated and enhanced through practice. The practice is best afforded through structured sequences of questions, preferably related to situations that are easily set up or that are commonly encountered in everyday experience. The specific examples given are not essential in themselves. They are merely illustrative, and many alternate versions and variations, of equal or greater effectiveness, can be generated by the individual teacher to best suit a particular group of students.

The algebraic results describing projectile motion, for example, should not simply be left for substitution of numerical values in end-of-chapter problems. Students should be led to interpret the algebraic results: What factors determine the range of the projectile? How does the total vector velocity change through the course of the flight? If a projectile is fired horizontally, what factors determine how far it has dropped below its initial level as the horizontal distance from the firing point increases? How can this vertical drop be decreased at a given range? What is the significance of the fact that the mass of the projectile does not appear in the equations?

Some years ago, on a Ph.D. qualifying examination, I asked why the cathode beam was not observed to be deflected by gravity when the electric and magnetic deflections were so pronounced. Upwards of 40% of the students taking the examination answered that this was due to the smallness of the electron mass, clearly indicating that they had not been led to engage in some of the thinking suggested above.

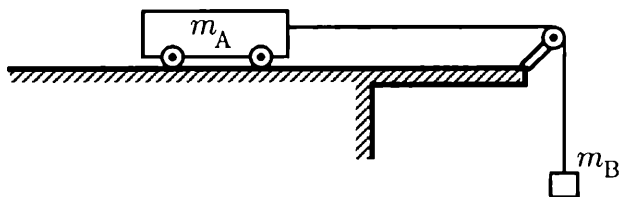


Figure 4.5.1 How does the force exerted by the string on the cart compare with the weight body B?

A situation that can be invoked to help students register a deeper understanding of Newton's second law is that of the ubiquitous problem, with zero friction and a massless string, illustrated in Fig. 4.5.1. This problem is usually posed for algebraic solution or numerical calculation (or both), but students

are rarely asked to bring out the essential physics by interpreting their results. They should be asked to extract what their algebraic results say would happen to the acceleration of the system and the tension in the string if m_B were made very much larger than m_A or very much smaller than m_A . Do these predictions make physical sense? Why or why not?

Finally, and most importantly, they should be asked how the force exerted by the string on the cart compares in magnitude with the weight of m_B , “comparing” meaning indicating whether equal, larger, or smaller. (I have, over the years, been asking this question of graduate students, most of whom are teaching assistants and have been helping students with this homework problem. Initially, virtually all of them have said that the two forces are equal in magnitude, and only about 30% changed their minds when I asked whether they would like to reconsider their answer.) Giving the response “equal” is an indication of inadequate comprehension of the phenomena in question and of the phenomenological implications of the Second Law. Being able to get correct answers to the conventional problems is no assurance of understanding.

A simple situation that affords a good exercise in phenomenological thinking and that does not lend itself to blind substitution in formulas is shown in Fig. 4.5.2: A person holds a block against a wall by exerting an inclined force F . There is friction between the block and the wall. Does the block move up, or down, or stand still?

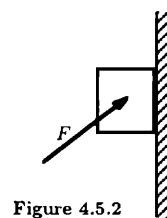


Figure 4.5.2

First, students should be led to draw the force diagrams for the block, the wall, and the hand holding the block, in accordance with the approach advocated in Section 3.16. The simplest version of the problem is then to give numerical values for the force, the mass of the block, and the coefficient of friction and to ask what happens in the given circumstances. The most sophisticated version is to ask for a discussion in algebraic terms alone. (The latter is an excellent exercise for students who are interested in theoretical work.) The point of this form of the question is that students are not told what to calculate; they must make some decisions as to what to do, what to look for, and how to interpret the results. This induces some phenomenological thinking and militates against reliance on memorized procedures.

The preceding problem might not be very interesting if it were an isolated situation unrelated to others to be encountered later, but this need not be the case. A valuable and closely related problem is that of the person “stuck” to the wall of the rotating cylinder in the amusement park. Another related problem is that of the electrically charged balloon sticking to the wall of the room. When students encounter these problems, weeks apart, as the course progresses, one begins to see the grins and glances that betoken recognition of an old friend in a new context. Were it not for such repetition or recycling, the ideas would be forgotten.

Finally, a two-dimensional situation rich in physical effects but rarely exploited as effectively as it might be is shown in Fig. 4.5.3. As usual, the force diagrams should be drawn first. Then the questions can begin.

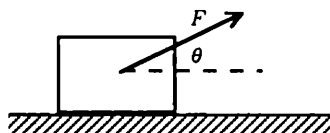


Figure 4.5.3

If the block is not sliding, is the frictional force necessarily zero? What range of values might the frictional force have? If the block is sliding, is the frictional force equal to μmg ? Why or why not? What happens to the acceleration as the angle θ is increased, starting at zero, while the magnitude of the force F is kept fixed? What happens if the angle is kept fixed and the magnitude of the force is increased indefinitely? What are some of the differences between the given situation and that in which the direction of force F is reversed?

4.6 RADIAN MEASURE AND π

Very few students in introductory physics courses have an understanding of radian measure, even if they have been exposed to the concept somewhere along the line. A few may have memorized something about arc length and radius, but they are rarely able to recover the definition precisely or to use it for dealing with angular displacement or angular velocity. For a great many students the problem begins not with radian measure itself but with the meaning of π .

If asked what π means, they are likely to mutter a formula, either for circumference or area, without being sure of which one the formula is for, but they are unable to explain what the symbol means or where the formulas came from. What is needed here, of course, is not a development of π to huge numbers of significant figures by rigorous methods of the calculus but evidence from direct measurement of circles—the primitive kind of evidence that must have been noted in ancient times. The calculation by methods of analysis can come later for those who continue to that level.

My own procedure with college students who do not understand the meaning of π is to have them use string and rulers to measure the circumference of every cylindrical object they can find around the laboratory and to measure the diameters by placing the objects between well-squared wooden blocks. They keep a running graph of circumference versus diameter. (It should be noted that this exercise is preceded by some of the graphing and arithmetical reasoning described in Sections 1.10 and 1.11, and the students have acquired some sense for the meaning of the slope of a straight-line graph.) Almost every time I have conducted such a session (with both pre- and in-service elementary teachers, for example) a voice has sounded through the room with something like “So that’s what they meant by π !!” This “discovery” must be followed by exercises and interpretation such as that illustrated in Chapter 1; otherwise the incipient understanding is lost.

Once the meaning of π is understood at the verbal level defined above, students can be led to see that the ratio S/R (where S denotes the length of intercepted arc and R denotes the radius) must have a fixed value for any given angle regardless of the size of R ; that the value runs from 0 to 6.28 for the complete circle; and that, because there is a unique value for each angle, the value of S/R can be used, in place of degrees, as a measure of the angle subtended. (Very few students in any introductory course I have taught, including calculus-physics, have known the meaning of the word “subtend.” If this term is used, one must be sure to define it explicitly.)

To enrich the context and add significance to S/R , it is worth pointing out that the ratios to which we give the names “sine,” “cosine,” and “tangent” also have fixed values for any given angle, regardless of the size of the circle on which the ratios are taken, and thus also measure the size of angles. These ratios, however, do not vary *linearly* with the size of angle (conceived of as a fraction of a complete circle) as does S/R , and they are therefore used in an entirely different way—another illustration of the importance of what is *not* the case.

Finally, the ratio S/R gives us a “natural” way of measuring angles by virtue of its connection to π , the intrinsic property of all circles; it is not “artificial” as is measurement in degrees. The principal property of the number given by S/R is that it is dimensionless; it is a “pure” number—this being part of the meaning of “natural” in this context. It should always be emphasized in this connection, however, that, although this new angular measure is *dimensionless*, it is not *unitless*: The unit is called “one radian” and can be carried around accordingly in calculations, but it must not be confused with dimensions such as mass, length, and time.

It should be noted that the preceding discussion follows the precept “idea first and name afterwards” advocated in Chapter 2. We thoroughly examine the properties of the ratio S/R before writing the expression $\theta \equiv S/R$ for subsequent use. This sequence of development helps reduce the tendency to memorize the formula without understanding that it is a definition, invented because of its practical utility.

Students who go on to levels at which they will encounter substitution of the value of the angle itself for the sine or tangent at small angles are helped by being led to confront the fact that the substitution is valid only in radian measure and not in degrees. This is very easily done with the now ubiquitous hand calculator, and most students enjoy making use of their calculators in this way. One way of presenting the problem is to suggest that they tabulate angles in both degrees and radians along with the sines and tangents, starting at larger angles and going toward smaller ones, and discover the angle at which the three values agree to two significant figures, three significant figures, four, and so on.

Finally, for students taking the calculus, it should be emphasized that the formulas for the derivatives of the sine and cosine functions (without inclusion

of a proportionality constant differing from unity) are valid only if the angle is measured in radians, since the limit relations

$$\lim_{\Delta\theta \rightarrow 0} \frac{\sin \Delta\theta}{\Delta\theta} = 1 \quad \text{and} \quad \lim_{\Delta\theta \rightarrow 0} \frac{1 - \cos \Delta\theta}{\Delta\theta} = 1$$

used in the derivations are valid only if the angle is measured in radians.

The principal reason that students previously exposed to radian measure have no understanding of the concept is not that it was not “explained” correctly (the explanations are usually basically sound); it is that they have never been required to discuss and explain any of the reasoning in their own words. Without such requirement, the tendency is to memorize formulas and avoid the visceral effort that accompanies striving for understanding.

4.7 ROTATIONAL KINEMATICS

The kinematics of circular motion in a plane is usually glossed over very quickly because of the obvious parallelism to rectilinear motion. For students who have genuinely mastered the concepts and relations of rectilinear kinematics, this is appropriate since unnecessary repetition would waste their time. As pointed out in Chapter 2, however, many students do not master the concepts on the first go-around, and some form of spiralling back is essential. The altered context makes a somewhat more careful treatment very worthwhile for this group, and the pace can be a bit more rapid than previously.

At this stage, polar coordinates are new to many students. Even if they have seen them in mathematics, they have not associated such coordinates with representation of physical events, and they should be given a chance to absorb the physical connections. They must see that angular positions, like positions on the number line, do not, in general, represent displacements of the moving object, which may never have occupied the zero position. They must have a chance to see how the algebraic signs arise for positive and negative angular displacements, for positive and negative angular velocities, and for positive and negative angular accelerations. The difficulties in discriminating between the various concepts (described in Chapter 2) arise over again for many students.

Another layer of difficulty is that of the arithmetical reasoning connecting angular displacement with linear displacement along the arc, angular velocity with tangential velocity, and angular acceleration with tangential acceleration. The reasoning involves not only an understanding of radian measure (Section 4.6) but also the perception that instantaneous tangential velocity corresponds in magnitude to the length of arc that would be swept out in one second if the angular velocity remained constant at the given instantaneous value. Many students must actually have the concrete experience of rolling a length of string off a cylinder in order to grasp the relations. While this is being done, it saves time to have them also examine the rolling of a wheel and to absorb the

connection between angular velocity and distance rolled in absence of slipping. The two contexts reinforce each other, and better understanding ensues.

After the kinematic relations have been developed, it is effective to demonstrate rectilinear and circular motions that correspond to each other. An especially useful concrete example is furnished by a wheel with a weight hanging from a string wound around the axle. If the weight is initially wound up to the axle, the wheel accelerates uniformly as the weight falls, and one can compare the parameters and the equations representing the observed unidirectional angular and rectilinear motions, respectively. If the weight starts near the floor and the wheel is given an initial spin, the wheel slows down as the weight rises, passes through zero instantaneous velocity, and reverses direction, speeding up as the weight falls. Many students do not perceive that the behavior of the wheel in these circumstances is identical with the behavior of a ball thrown vertically upward until they are led to articulate the connection.

As pointed out in Sects. 1.6 to 1.10, many students in introductory physics courses start out with serious difficulties stemming from lack of mastery of arithmetical reasoning with ratios and division. Those who have not acquired such mastery by the time they arrive at rotational kinematics have great trouble with the simple connections among angular velocity, period of revolution, and frequency of revolution. If not required to explain the algebraic relations in their own words, they memorize the relations without understanding and are unable to use them in a sequence of reasoning. In solving problems, they mechanically substitute numbers in formulas, hoping to arrive at the answer given at the back of the book. Having no other frame of reference, they consider the task accomplished when the “answer” emerges, and they are unaware that they have no understanding of the analysis. For such students, encounter with the simple rotational quantities is a valuable opportunity to spiral back, in altered context, and strengthen their facility with ratio reasoning.

4.8 PRECONCEPTIONS REGARDING CIRCULAR MOTION

Many students come to the study of dynamics of a particle in circular motion with a number of very deeply rooted preconceptions—preconceptions that tenaciously persist beyond earlier science or physics courses where the subject might have been considered. One such preconception is the expectation that curvilinear motion imparted to an object by some constraint persists after the constraint is removed. For example, when shown the apparatus in Fig. 4.8.1 in which a ball is set in motion around the inside of a open hoop lying on a table, and asked what will happen when the ball reaches the open section, many students expect the ball to follow a curved path at the opening and to continue around the inside of the hoop. One hears gasps of astonishment when the demonstration is performed, and the ball leaves the hoop on the tangent

line without continuing around.

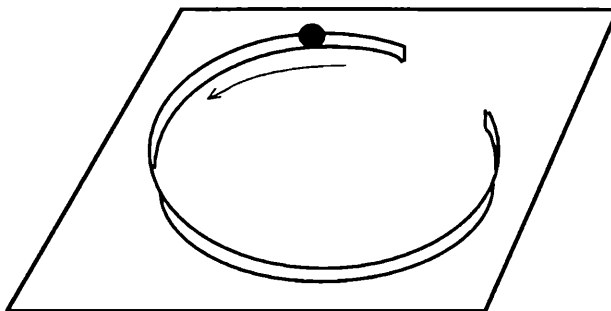


Figure 4.8.1 Ball rolling on table inside an open hoop does not continue on curved path inside hoop.

McCloskey and his co-workers [McCloskey et al, (1980); McCloskey (1983)] report similar results in interviews with students who are asked to predict the exit path of a bead fired around inside a spiral glass tube lying on the table. Many students predict the exit path to be curved in the same direction that the tube is curved; there is a strong expectation of persistence in a curved path after the constraint is removed.

There is much to be said for starting study of the dynamics of circular motion of a particle with at least some of the following demonstrations: (1) the hands-on imparting of approximately circular motion to the large block of dry ice on a glass plate as suggested in paragraph 8 of Section 3.10 (if at all possible); (2) the demonstration of Fig. 4.8.1; (3) cutting the string holding a puck in circular motion on an air table (or holding a dry ice puck in circular motion on a glass plate); (4) the curved tube experiments used by McCloskey.

Students should be invited to predict what will happen in each case—and to argue about their predictions—before the demonstration is performed. Such procedure is far more effective than prior assertion of the expected result by the teacher.

A second deeply rooted preconception is that of the presence of a force in the direction of motion: The push imparted to the object “stays with the object and keeps it moving”; the bob or puck lags behind the string, and the string is pulling it in the direction of motion.² One could define the “Compleat Optimist” as the teacher who expects all the students to have shed this idea

²Many students, when they whirl a bob on a string, believe they are pulling the bob in the tangential direction. Since the application of torque accelerating the bob in the first place, and then keeping it going in the face of frictional resistance, is actually a rather complicated process, this view should not be surprising. One should not jump to the idealization of negligible torque as though it were obvious to everyone. The notion that the string keeps the bob moving by pulling in the tangential direction provides an additional reason for doing experiments with an inward force supplied by a solid body, such as the hoop in Fig. 4.8.1; it is more difficult to rationalize the presence of a tangential force in this situation.

after lucid treatment of the law of inertia for rectilinear motion along lines such as those of Chapter 3. One should not be surprised, however, that the expectation is not realized; the yield, as in chemical reactions, is never 100 percent. The context has been altered from rectilinear to circular motion, and, although some students have indeed mastered the Newtonian view, there remain significant numbers who have not. The altered context should be seized upon as an opportunity to spiral back and bring additional recruits to the Newtonian level.

A third, and very widely prevalent, preconception concerning circular motion is that the object is being “thrown outward.” A principal source of this (apart from hearing it said by poorly prepared teachers in poorly taught science courses) is one’s own sensation in going around curves. It is for this reason that it is very advantageous to start laying the groundwork to counter this idea back in the discussion of rectilinear motion as suggested in Section 3.20. Once one begins to realize that he or she is not thrown backward in a car that is speeding up, or forward in a car that is slowing down, one is far better prepared to discern what is happening in going around a curve, and the way is then better prepared for formation of the concept of centripetal force.

4.9 CENTRIPETAL FORCE EXERTED BY COLINEAR FORCES

Since the concept of centripetal force and the relation

$$F_{\text{cent}} = \frac{mv^2}{R} \quad (4.9.1)$$

are most commonly developed in connection with a bob (or frictionless puck) attached to a string and moving in a horizontal circle, there is a strong tendency for many students henceforth to connect the term “centripetal force” with the pull of a string regardless of whatever other effects might be present simultaneously.

Although it might seem trivial to the teacher, many students are helped if one emphasizes the absence of a string in the case of the ball rolling around the inside of the hoop (Fig. 4.8.1) and in the case of our own body in the car going around a curve on a level road. Students need to be made explicitly conscious of the fact that a centripetal effect might be a push toward the center rather than a pull.

A valuable problem at this point is that of the car on the unbanked road. It offers the opportunity to calculate centripetal forces that are not exerted by strings: The centripetal force exerted by the seat on one’s own body and the frictional force exerted by the road surface on the car. (This problem, stopping at this point, paves the way for subsequent treatment of banking of the road, the intervening time helping in the assimilation of the ideas.)

An element of physical insight rarely made explicit for the student is the fact that in these various circumstances we are invoking a passive force (exerted by the string, or hoop, or seat, or road surface) that adjusts itself to the magnitude required by the given radius and angular velocity just as the normal force exerted by the table on a block adjusts itself to the magnitude required by the weight of the block, or the frictional force exerted by the floor on a block (before sliding) adjusts itself to the external push or pull on the block. Unless led to articulate this insight, many students fail to see that the string will break, that the hoop will bend, and that the car will slide when the demand for a still larger force can no longer be sustained. To see this connection among these disparate situations is highly conducive to learning.

Situations depending on passive forces must eventually be set in contrast with those involving active forces such as gravity or electrostatic attraction. In these latter cases, a change in tangential velocity is not compensated by a change in the passive force, and the radius must change in consequence. Many students do not penetrate to an understanding of the physics of these contrasting situations unless they have the opportunity to discuss the differences. Such discussion is frequently omitted because treatment of the active force cases comes quite some time after the initial development of circular motion, and advantage is not taken of the opportunity to spiral back. (No pun intended.)

Furthermore, because situations other than the one represented by the equality in Eq. 4.9.1—that is, with forces larger or smaller than the value demanded in the equation—are almost never examined, students acquire the impression that any inwardly directed force is to be called a “centripetal force.” This impression needs to be countered by asking how the bob will behave, at a given radius and tangential velocity, if the inward force is made larger than mv^2/R or smaller than mv^2/R . It is not at all necessary to work out the consequences mathematically; this is far too formidable a task. All one needs is the qualitative insight that the particle will deviate either inward or outward from the prescribed circle. When the student has been led to see the broader perspective (i.e., what is *not* the case as well as what is), the operational definition of “centripetal force” begins to take on firmer meaning: The term applies only to the magnitude of the force that imparts the particular acceleration mv^2/R to the particle of mass m and keeps it in the given circular path of radius R .

A still more serious problem about the meaning of “centripetal force” emerges in circumstances where the centripetal force is imparted by two or more effects—as in the case of a bob on a string revolving in a vertical circle or in the case of the car on the loop-the-loop. Although it should be emphasized from the very beginning that the term “centripetal force” refers always to the *net* force that imparts the acceleration v^2/R to the object in question, the word “net” means very little to students until they encounter the case of superposed forces. With the bob at the top or bottom of its circle, they first

tend to regard the centripetal force as being just the pull of the string and do not recognize that, in this case, the term refers to the algebraic sum of the pull of the string and the weight of the bob. The analogous difficulty arises in the case of the loop-the-loop.

This misconception regarding the meaning of the term “centripetal force” then impedes understanding of the physical phenomenon—namely, that the active force (the weight of the bob or car), contributing to the centripetal force, remains unchanged, while the passive force (pull of the string or push of the track) adjusts itself to precisely the value that makes the algebraic sum of the two forces impart the centripetal acceleration required.

Many students solve the end-of-chapter problems that ask for the smallest tangential velocity that allows the bob or car to negotiate the top of the circle (or the height at which the car must start on its initial track) by memorizing an established procedure and emerge with no understanding of the role of the two superposed forces or of the full meaning of the term “centripetal force.”

The weakness of these conventional problems is that they concentrate all attention on the crossover condition and fail to lead the student to examine what must happen, in general, on either side of the crossover: What is the magnitude of the tension in the string at the top of the circle if the tangential velocity of the bob is *greater* than the critical value? What happens to the tension as the tangential velocity is decreased? Under what circumstances does the tension become zero? What happens if the tangential velocity is decreased further? The whole sequence must be analyzed and interpreted if understanding is to be cultivated.

The most effective way of doing this is to lead the students to set up the equation for tension in the string at the top of the circle:

$$T = \frac{mv^2}{R} - mg \quad (4.9.2)$$

and examine how T changes as one causes v to decrease from some initially high value. They then encounter both the crossover condition and the necessity of interpreting the meaning in the change of sign of T at the critical value of v . This is a valuable exercise in interpreting mathematical results. A parallel analysis can then be invoked to contrast the variation of T at the bottom of the circle with its variation at the top.

Throughout these discussions it should be emphasized that T is *not* the centripetal force; that the centripetal force at the top of the circle is given by $T + mg$ and at the bottom of the circle by $T - mg$. This begins to eliminate the notion that T is always the centripetal force.

Although it might now appear that examination of the normal force exerted by the track on the car in the loop-the-loop has become a trivial repetition, this is actually not so. The context is just sufficiently altered so that analysis of the variation of the normal force exerted by the track at the top and bottom of the loop makes a good test question. Students who have mastered the insights

enjoy doing well, while those still struggling have a valuable re-encounter. Furthermore, there is a connection between this situation and the elevator sequence developed in question 10 in Section 3.25: The sensation experienced by the loop-the-loop rider (at the bottom of the loop) is similar to the sensation imparted by an elevator in upward acceleration.

Still another useful variation is the case of a car going over the crown of a hill, the crown being treated as essentially circular in shape. Here one finds oneself on the outside of a “track” instead of on the inside (as in the loop-the-loop). Again, the situation is slightly different from that of the preceding cases, and the change is just enough to be a sensitive probe for understanding. This situation ties in with a sensation many students have experienced: That of seeming about to “take off” when speeding over the top of a hill. It also connects with the sensation in an elevator accelerating downward. Tying all these cases together in homework provides a valuable learning experience.

Relatively few students acquire firm mastery of force diagrams during the initial exposure to rectilinear motion, and it is advisable to utilize all available force problems in circular motion as an opportunity to spiral back and strengthen grasp of the force concept. For this purpose students should be required, on both tests and homework, to draw the force diagrams, describe forces in words, and identify third law pairs in the manner recommended in Section 3.12. This means that the diagrams should include not only the moving particle but also the string and whatever is at the other end of the string; the track as well as the car in the loop-the-loop; one’s own body in the car, the car, and the road surface when going around a curve on an unbanked road.

4.10 CENTRIPETAL FORCE EXERTED BY NONCOLINEAR FORCES

In the following illustrations, students should, of course, be required to draw force diagrams for all of the interacting objects in the manner outlined in Section 3.12.

1 *Banking of a road.* This problem is usually treated in a manner similar to the treatment of the car negotiating the loop-the-loop. Only the crossover condition is solved for, and many students fail to achieve understanding of the physics. In this instance, it is particularly important to start with an understanding of what is involved in negotiating a curve on an unbanked road: The location of the center of the circular motion, the role of friction, and the origin of skidding. This is why treatment of this case is so strongly recommended in the preceding section.

After the unbanked case is understood, one can lead students to pursue what happens as the angle of banking is increased from zero. First, however, they must be led to define an appropriate set of coordinates. Because virtually all inclined plane problems previously encountered have been most

conveniently handled with Cartesian coordinates parallel and perpendicular to the plane, most students have developed the conditioned reflex that one always chooses this orientation of the axes. They must be brought to the realization that, in this instance, it is better to take one of the axes in the direction of the required centripetal force (i.e., horizontally), else one invites a great deal of misery.

Having established the most appropriate coordinate axes, students should then be led to the qualitative insight that, as the angle of banking increases and the normal force exerted by the road surface on the car develops an increasing horizontal component, the magnitude of the required frictional force decreases. The optimum condition then becomes clearly defined and can be solved for. The analysis should not stop here, however; it should continue into what happens when the critical angle is exceeded, and students should articulate the renewed demand on the frictional force, now acting outward rather than inward.

2 *Amusement park ride in which one can take one's feet off the floor while remaining stuck to the outer wall of a rotating cylinder.*

This problem affords a very useful spiralling back to connect with situations in which a box was pressed against a vertical wall by a horizontal force and did not slide down. The two problems should now be juxtaposed and examined for similarities and differences in the physics. The rotating case should be examined in full, setting up the appropriate equations, imagining continuous change in angular velocity (starting at either a high or a low value), passing through the critical value, and continuing on to the other side. This exercise is especially valuable for students going on to more advanced levels because it requires thinking in terms of inequalities more than in terms of equalities—a kind of thinking they very rarely have the opportunity to do in elementary physics.

3 *Direction of g (the local vertical) on a spherical (i.e., rigid), but rotating earth.* (Although this is not an especially formidable problem, it offers considerable difficulty to most students, even honors students in a calculus-physics course, and it therefore absorbs a substantial amount of time. My own practice is to toss it out as an extra-credit challenge in ordinary classes but to assign and discuss it in honors sections of students going on to higher levels.)

This highly idealized problem can first be examined purely qualitatively with great conceptual profit; a more quantitative treatment can come later, if at all. In my own experience, the most effective way is to establish a diagram of the rotating earth such as that in Fig. 4.10.1 and imagine suspending a plumb bob on a string. Students must first be led to recognize that the centers of circular motion at different latitudes lie on the intersection of the plane of the latitude circle and the axis of rotation rather than coinciding at the center of the sphere. For most students, this is a first encounter with the rotating

sphere, and they are not yet familiar with this simple aspect; the encounter is useful since it prepares them for more complicated cases later.

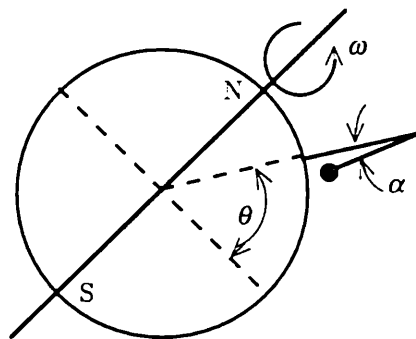


Figure 4.10.1 Plumb bob suspended above a perfectly spherical, rotating earth.

The students are then led to struggle with the vector diagram of the forces acting on the plumb bob: They must recognize that the gravitational force is fixed in magnitude and directed toward the center of the sphere; they must then discern that the passive force exerted by the string must adjust itself in magnitude and direction in such a way that the resultant of the two forces provides the requisite centripetal force directed along the perpendicular to the axis of rotation. The force exerted by the string is, of course, equal and opposite to the “local weight” of the bob, and its direction defines the local vertical. It is then apparent that g is directed toward the center of the earth at both the equator and the poles and that, at other latitudes, there is a deviation in the direction shown in Fig. 4.10.1. From these results students can begin to see how the figure of the oblate, non-rigid earth would be formed, and one can visualize what would happen with increasing and decreasing angular velocity ω .

It should be noted that the introduction of the string is a very useful device. It is much easier to deal concretely with this situation than to try to visualize what happens with a freely falling body.

Once the vector diagram has been drawn, a quantitative solution is virtually at hand. Neglecting small terms, the result is

$$\alpha \cong \left(\frac{a\omega^2}{g} \right) \sin 2\theta \quad (4.10.1)$$

where α denotes the angle of deviation of local vertical from the radial direction to the center of the sphere, a the radius and ω the angular velocity of the earth, and θ the angle of latitude. It is apparent that the angle of deviation is largest at a latitude of 45° . The magnitude of this maximum deviation is close to six minutes of arc.

4.11 FRAMES OF REFERENCE AND FICTITIOUS FORCES

Most teachers are concerned about how to deal with the concept of “centrifugal force,” since students come to physics courses with this vocabulary firmly embedded through frequent encounter in school science and in popular literature. The usage is also massively reinforced by the personal sensation of being “thrown outward” when going around a curve. It seems to me that the most appropriate treatment has been clear for a long time and is well handled in many texts. An especially fine treatment is available in the *PSSC* film “Frames of Reference,” narrated by Hume and Ivey of the University of Toronto. (I have long felt that this film can stake a legitimate claim to being the best instructional physics film ever made. The physics is excellent. The film is not badly dated even though it is in black and white, and its humor stands up under repeated viewing.) The approach is to define what is meant by a “fictitious force” and identify “centrifugal force” as being in that category.

If one does not wish to get involved in the subtlety of fictitious forces (and such a decision is perfectly legitimate, especially in terminal courses), the best approach is to deny the validity of centrifugal force as a concept applicable in the inertial frames of reference in which we are applying the Newtonian laws and to eliminate the use of the term. As indicated both in Chapter 3 and in preceding sections of this chapter, it can be compellingly argued that there is no force throwing an object outward in circular motion and that only a net inward force is necessary to impart centripetal acceleration.

For students going on to more advanced levels, however, it is worth developing the idea of the fictitious force since it will, of necessity, be encountered in non-inertial frames of reference, especially the frame of the rotating earth itself in subjects such as meteorology, oceanography, or geophysics. Here the important fictitious force will be the Coriolis force rather than the centrifugal force, but it is wise to start with the simpler concept.

Some texts and teachers elect to accept the term “centrifugal force” and apply it to the force exerted by the bob on the string or by the string on the peg to which it is tied. This is a most unfortunate and undesirable way of treating the concept. The unwritten text that goes with the term “centrifugal force” is that this is a force acting *on the object* in circular motion and *not* a force exerted by the object on something else. Such locutions help perpetuate confusion about third law pairs. Furthermore, the preconception held by the learner is that of an outward force on the object itself and not that of a force the object exerts on its surroundings. A force acting on the object itself is legitimate only as a fictitious force in the noninertial frame. There is no point saddling students with a specious version that they will have to change when they get to the more advanced subject matter.

4.12 REVOLUTION AROUND THE CENTER OF MASS: THE TWO-BODY PROBLEM

Many treatments of circular motion, especially in algebra-based courses, confine themselves to the case of a fixed center. This is proper for very elementary physical situations such as those discussed in the preceding sections but, since allusions are usually also made to the revolution of the moon around the earth and of the planets around the sun, it leaves students with the impression that the latter motions are identical with the former. Furthermore, any mention of lunar and solar tides is specious without some accompanying consideration of the two-body problem.

I have never had a student ask, for example, why we should view the earth as revolving around the sun when the gravitational interaction, as a centripetal force, might just as well make the sun revolve around the earth? There is valuable and important physical understanding and phenomenology buried here, and the question should be elicited since it does not arise spontaneously.

This need not be done with full dynamical analysis, including formation of the concept of reduced mass, although such treatment is well within the scope of the usual calculus-physics course. It is possible to handle the question entirely qualitatively with students unprepared for an analytical approach.

I have, for example, used the following approach with classes of general education students and elementary school teachers: Early in the course [Arons 1977], I pose the question as to why we believe the earth and planets revolve around the sun rather than vice versa, and we spend weeks working up toward an answer to the question. After having developed kinematics, Newton's laws, the concept of terrestrial gravity, and the concepts of centripetal acceleration and centripetal force, we read about Newton's suggestion that gravity might extend to all the interactions of the celestial domain and that it might provide the necessary central force for the observed motions. The question remains as to what revolves around what.

Some days after having raised this question and allowed it to simmer, I go to the air table and tie two identical pucks together with a string. I tell the class that I am going to give the pucks a push such that they will tend to go into circular motion with as little translation as possible, and I ask for a prediction of the location of the center of the circular motion. The first time I did this I expected a wide range of response, especially because I hinted that I would hold one puck still and push the other. To my surprise there was a chorus of expectation of circular motion around the midpoint between the two pucks. (Our intuition for symmetries of this kind is apparently very deeply embedded and extends even to individuals without much prior experience with the phenomena.)

Having established the behavior of the identical pucks, I proceed to load one of the pucks, and I ask again for a prediction of the center of circular motion. The chorus places the center closer to the more massive puck. When

I ask what will happen as the mass increases further and becomes very much larger than that of the unaltered puck, the expectation of a center of rotation very close to the center of the massive puck emerges very clearly. Almost always someone in class exclaims, usually in the startled tone that accompanies discovery, "That must be the earth and the sun!" We are now on our way; there remains only the problem of developing evidence as to relative masses.

If one has a bit more time, with a somewhat more sophisticated class, one can set down and interpret the equation

$$mr\omega^2 = MR\omega^2 \quad (4.12.1)$$

for the two-body case and bring out the role of the center of mass of the system. Although this is desirable, it is not essential as a first-level approach.

If one essays to discuss the lunar tides, however, other than by pure hand waving, it is essential to develop the significance of revolution of both the earth and the moon around the center of mass of the system. If the earth were fixed, with the moon revolving around it, there would be only a diurnal (rather than a semidiurnal) lunar tide, and its height would be devastating. The semidiurnal tide and its modest height stem from the fact that the earth is in free fall toward the center of mass around which the revolution is taking place.

One way of taking up this issue is to extend the ideas developed in the gedanken experiments in a freely falling elevator in question 10 of Sect. 3.25. (This is another valuable opportunity for spiralling back.) Suppose we make the height of the elevator *very* large—so large that it reveals the gradient of the earth's gravitational force, with the force being measurably smaller at the top of the elevator than at the bottom.

In free fall, the entire body of the elevator accelerates downward with the acceleration at the center of gravity; let us denote the magnitude of this acceleration by g_{cg} (The elevator itself is under stress since the gravitational force at the bottom is greater than that at the top.) If we release a ball in the vicinity of the center of gravity, it will stay where we release it. If we release a ball at the top of the elevator, however, it will "fall" *upward* relative to the elevator with an acceleration of magnitude $g_{cg} - g_t$, where g_t denotes the free fall acceleration relative to the earth at that upper elevation. Similarly, if we release a ball near the bottom of the elevator, it will accelerate "downward" relative to the elevator with acceleration $g_b - g_{cg}$. These upward and downward accelerations would be noticeable only because the gravitational effect of the elevator itself would be negligible.

If the elevator had a powerful gravitational field of its own originating at its center, this gravitational force would completely overpower the effects described in the preceding paragraph, and the ball would fall toward the center of the elevator whether it were released at the top or the bottom.

The situation of the earth in the earth-moon system is to some degree analogous to that of the freely falling elevator: The earth is in free fall toward

the center of mass of the system; the acceleration of the entire earth is that of its center gravity; the earth is subjected to the gradient of the lunar gravitational force and is under stress. Relative to the earth, there is now a residual (fictitious) force away from the moon on the far side of the earth and toward the moon on the near side. This “splitting” accounts for the semidiurnal lunar tide.

At this point the analogy to the elevator begins to break down, and one must be careful. The stresses produced by the residual lunar forces result in earth tides in the solid earth; although measurable, these are very small. Our interest is in the oceanic tide. Here the sphericity of the earth becomes a critical factor. We resolve the residual lunar forces into *radial* and *tangential* components with respect to the *earth*. The radial component is completely overpowered by the earth’s own gravity and has a negligible effect (as in the case of the freely falling elevator with a powerful gravitational field of its own.) It is the *tangential* component that is capable of displacing ocean water until the hydrostatic pressure gradient due to the slope of the surface balances the tangential effect of the residual lunar gravity on either side of the earth.

Many textbook discussions of the semidiurnal lunar tide tend to leave the impression that it is the residual *radial* force directly “underneath” the moon (and toward the moon) on the near side, and the residual fictitious force radially away from the moon on the far side, that pile up the waters and produce the semidiurnal tides. This is entirely incorrect, as indicated above: it is the *tangential* force that dominates the observed effects.

Lamb (1932) presents a rather neat alternative view of the semidiurnal lunar tide. He shows that, if there were *two* identical moons, each with half the mass of the existing moon, orbiting the earth 180° apart, the earth would be fixed rather than in free fall, but the tidal situation would be identical with the one that exists. [More detailed discussion of all these matters can be found in Tsantes (1974), Arons (1979), and Lamb (1932), as well as in additional references cited in these sources.]

A final word about the two-body problem: It arises again, in a highly significant way, in the quantum theory of the hydrogen atom. To obtain agreement between the theoretically calculated and empirically observed Rydberg constant, one must invoke the reduced mass of the electron-proton system rather than just the mass of the electron. The latter usage implies a fixed proton while the former acknowledges dynamics relative to the center of mass.

An interesting physical insight attends awareness of the difference between the earth-moon and electron-proton systems. If the earth were fixed, the period of revolution of the moon would be given by the (Kepler’s third law) form

$$T_1^2 = \left(\frac{4\pi^2}{Gm_E} \right) H^3 \quad (4.12.2)$$

while the period of revolution about the common center of mass is given by

$$T_1^2 = \left[\frac{4\pi^2}{G(m_E + m_M)} \right] H^3 \quad (4.12.3)$$

where H denotes the distance between centers of the earth and moon. Thus the actual sidereal period of the system is slightly less than it would be if the earth were fixed.

Equation 4.12.3 does not contain the full expression for the reduced mass because of the cancelation arising from the mass dependence of the law of gravitation. The full expression for reduced mass does appear, however, in the electron-proton case because the force law (Coulomb's law) does not involve the masses. Many students coming to the hydrogen atom problem with no previous inkling of the phenomena attending the two-body situation, are confused and astonished by the unfamiliar physics.

The two-body problem enters in a significant conceptual way in a somewhat different gravitational case—that in which a space vehicle acquires additional speed and kinetic energy by “swinging around” a planet. In the frame of reference of a stationary planet, a space vehicle, entering and leaving on a hyperbolic orbit, could not acquire additional speed; it would exit with the same speed at which it entered.

In the frame of reference of the solar system, however, a space ship, in swinging around a moving planet in a properly chosen orbit, acquires additional speed and kinetic energy from the planet's motion. This device has been employed on numerous space missions, for example in the “Venus-earth-earth gravitational assist” that made it possible for the Galileo vehicle to reach Jupiter.

Given these various instances, it is apparent that students (especially those going on to more advanced subject matter) can be significantly helped by at least a qualitative exposure to two-body physics in the introductory course.

4.13 TORQUE

Once the preceding conceptual structure is established, it may seem that the concept of “torque” is a simple extension that requires relatively little effort and attention. For many students, however, this is not the case. They are still struggling to grasp the large array of new concepts, the analytical formulations, and the connection to phenomena. In many instances, the expression for torque is asserted much too quickly, with far too little motivation or connection with experience. Furthermore, the effectiveness of development is enhanced if one adheres to the precept “idea first and name afterwards.”

In addition to appealing to kinesthetic experience in the acceleration of rotation with cranks and various lengths of crank arms, it frequently helps to draw on students' previous experience with balancing (as in the see-saw).

Lessons on balancing are quite common and widespread in school science, and invoking concrete experience from the past does much to reduce the feeling of fear and insecurity that attends new abstract formulations.

If students have had previous exposure to beam balancing along the lines of the prototype illustrated in Fig. 4.13.1, it is likely that they have described the condition for balance in terms of ratios of the form

$$\frac{W_1}{W_2} = \frac{L_2}{L_1} \quad \text{or} \quad \frac{L_1}{W_2} = \frac{L_2}{W_1} \tag{4.13.1}$$

(They may not have written out the expressions algebraically, but their thinking has usually been in this mode, principally with integer ratios.)

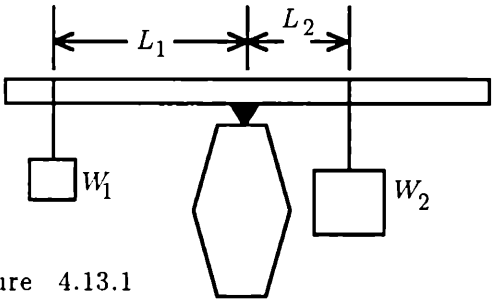


Figure 4.13.1

One can draw on this background in the following manner: Set up a situation such as that illustrated in Fig. 4.13.1; elicit from the students a description of the balance condition, and also elicit both algebraic statements in Eqs. 4.13.1. Then point out that the balance condition in the form of ratios “scrambles” the forces and the lever arms, that is, the subscripts 1 and 2 both appear on each side of the equation. Now ask whether one can restate the balance condition so that all the subscript 1’s are on one side of an equation and all the 2’s are on the other and elicit the form

$$W_1 L_1 = W_2 L_2 \tag{4.13.2}$$

It can then be pointed out that Eq. 4.13.2 states the balance condition in such a way that the left side of the equation contains only terms from the left side of the balance and the right side contains only terms from the right side of the balance. Thus the quantities $W_1 L_1$ and $W_2 L_2$ can each be regarded (or interpreted) as “effects” intrinsic to each side, and that, even more specifically, they can be interpreted as measuring a “turning effect” applied to each side. This interpretation is strongly reinforced by showing that the beam is unbalanced and rotated counterclockwise if $W_1 L_1$ is made greater than $W_2 L_2$, and is unbalanced and rotated clockwise if the inequality is reversed. (It is important to invoke departure from the balance condition; otherwise many students fail to see what is *not* the case in addition to what *is*).

Next, it is useful to set up a situation with two weights on one side of the fulcrum and one on the other as shown in Fig. 4.13.2. Now students can be

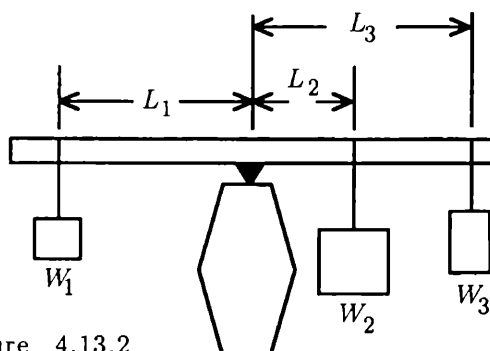


Figure 4.13.2

led to see that the condition for balance can no longer be expressed in a form such as that of Eqs. 4.13.1 but that the form of Eq. 4.13.2 still works. The equilibrium condition becomes

$$W_1 L_1 = W_2 L_2 + W_3 L_3 \quad (4.13.3)$$

This strongly reinforces the interpretation of the WL product as a “turning effect,” and the experimental observation shows the effect to be simply additive! One can now proceed to introduce the concept of the “lever arm” of an applied force.

Having established the significance and utility of the WL product, one can now examine a few situations in which pushes and pulls are applied to the beam by a force other than a hanging weight (pull on the string by the hand, for example) and point out that we have been dealing, initially, with a very special situation—one in which the applied force has been perpendicular to the lever arm. How should the turning effect be calculated when the force and lever arm are not mutually perpendicular?

This is an opportunity to spiral back, in a new context, to the concept of force components. One can resolve the applied force into components along, and perpendicular to, the lever arm and examine the effect of each component. Since the component along the lever arm is directed through the turning point, it has zero turning effect. The turning effect is determined by the size of the component perpendicular to the lever arm, and so on. This is a valuable opportunity to reinforce awareness that vectors are not anchored to points but are to be shifted, in our imagination, along their lines of action. It is also very important to show the alternative calculation of the turning effect in which one finds the effective lever arm for the total applied force, without resolving the latter into components.

Many students have very great initial difficulty in making these calculations. The geometry is far from transparent to them, especially when constructions have to be added to an initial diagram. Much more practice is needed than is usually afforded in end-of-chapter problems. Practice should

include a wide variety of configurations (rotated into different orientations when represented on a page) and the labeling of different angles in the different configurations. Problems should include zero length lever arms as well as lever arms that must be visualized and are not concretely represented in the given drawing, and problems should include algebraic formulations as well as numerical calculations. This is another case in which practice could be very effectively implemented through well-designed, computer-based modules.

After the concept of “turning effect” has been firmly registered along lines such as those illustrated above, it becomes appropriate to introduce the name “torque.” Such concrete and relatively modest development paves the way for subsequent treatment of torque as a vector cross product and as the agent of change of angular momentum, and it makes the latter abstract extensions far more intelligible to many students.

4.14 SAMPLE HOMEWORK AND TEST QUESTIONS

NOTE: The following qualitative questions on force diagrams illustrate the spiralling back that is helpful to many students. They review ideas developed at the level discussed in Chapter 3 and combine them with ideas discussed in this chapter. Despite the previous exercises, many students will still have difficulty with these questions when they are first posed. They are helped by the practice. Other similar questions, connected to actual or possible personal experience, are needed. The interaction of permanent magnets might well be included, especially if magnets were used to show noncontact interaction between gliders on an air track.

1 A person sits on a box resting on the platform of a merry-go-round and holds a bob suspended on a string. The merry-go-round is turning. Draw well-separated force diagrams of the bob, the string, the person, the box, and the platform. Show forces of equal magnitude with arrows of equal length; show larger forces with longer arrows, etc. Describe each force in words, and identify the third law pairs.

2 A person stands at the edge of a rotating merry-go-round and leans in such a way as to feel most comfortable and well balanced during the rotation. Hanging from the railing near the person is a pendulum (a bob on a string.) It is an observed fact that, under these conditions, the body of the person is aligned parallel to the string of the pendulum. Draw force diagrams of the bob, the string, the railing, the person, and the platform. Explain how it comes about that the person is most comfortable under the conditions described. Describe the connection between this situation and the situation of a car going around a curve at the speed for which the curve is banked. A person sits inside the car, and a pendulum bob is suspended from the roof.

Chapter 5

Momentum and Energy

5.1 INTRODUCTION

There are several sources of difficulty for students entering study of the momentum and energy concepts: (1) The necessity of defining the system that will be under consideration; (2) distinguishing clearly between open and closed systems; (3) mastering the large number of new technical terms that are generated—many of them drawn from everyday speech but endowed with altered meaning—by connecting the new technical terms with the algebraic expressions and calculations with which they are associated; (4) mastering the relations among the technical terms and using them to describe and analyze familiar, everyday phenomena.

These difficulties can be significantly reduced if students arrive at this level with prior practice in: (1) Stating operational definitions in their own words; (2) translating symbols into words and words into symbols; (3) describing familiar phenomena in the technical vocabulary previously developed.

If the material on kinematics and the laws of motion has been covered too rapidly, however, without opportunity for the kind of practice just referred to, the difficulties with the abstractions now encountered are seriously compounded, and many students fall by the wayside. This is not to say that the previous material must have been fully *mastered*; the exposure must have been such that the student is ready to progress into the new abstractions by using procedures that are becoming familiar through encounter in earlier, more concrete, contexts. Concepts not fully mastered previously will become more firmly assimilated through the spiralling back afforded in the new, richer context.

A serious impediment to understanding of the energy concepts, and of the concept of “work” in particular, resides in the way many textbooks introduce these ideas. The concept of “work” and the work-kinetic energy theorem are initially developed for the very special case of point masses or particles, that is, the case in which the displacement of the net external force is equal to the displacement of the center of mass of the object on which the force acts. The

indicated relations are then extended, sometimes fallaciously, to systems (such as rotating or deformable systems) with internal degrees of freedom in which the displacement of the external forces is *not* equal to the displacement of the center of mass. The vocabulary becomes very confusing; many commonly made statements are basically incorrect. Many of the better students are very uneasy and feel that something is missing, but they fail to pursue the issue for fear of appearing “stupid;” weaker students simply memorize without comprehension. This problem requires more careful attention in texts, lectures, and homework; it is discussed in Sects. 5.6 to 5.14.

5.2 DEVELOPING THE VOCABULARY

Having become very used to the concepts over an extended period of time, teachers and textbook authors tend to lose sight of how bewildering the large array of new technical vocabulary is to many students. Confusion is also augmented by the fact that new meaning is being given to seemingly familiar words such as “momentum,” “energy,” “work,” and “impulse.” Besides the individual nouns themselves, we have phrases such as “impulse delivered (or work done) by the force P ”; “impulse delivered (or work done) by the net force F_{net} ”; “*change* of momentum of body A”; “*change* in kinetic energy of body B”; “work done *against* force P or force f ”; “change in potential energy of system S.”

Furthermore, in rectilinear situations, the algebraic sign associated with change in momentum of a body has very different meaning from the algebraic sign arising in connection with change in its kinetic energy.

For most students, whether in algebra- or calculus-based courses, reading text descriptions, hearing the lecturer apply the vocabulary to a few sample cases, and doing the end-of-chapter problems do not provide sufficient exercise in use of the vocabulary. Understanding the meaning of the terms and phrases, and the ability to use them in describing physical phenomena, do not develop until students have had the opportunity to establish sequences of interconnection between words and numbers in an array of simple physical situations. Furthermore, there is a good bit of memorizing to be done, and students fail to recognize the necessity of memorizing vocabulary unless such necessity is explicitly pointed out and acquisition is tested for.

A few sample problems that help develop this capacity are given in the last section (5.16) of this chapter. The reader will note that the individual segments of these problems are to be found in most textbooks. The gap left by the textbooks resides in the fact that the segments are presented separately and are rarely linked into one integrated sequence, applying the vocabulary to a rich, extended context. Any teacher who has never given, on a test, a simple sequential problem such as one of those illustrated in Sect. 5.16 will be shocked at how poor the entire class performance is the first time this is

done. The performance is extremely poor even in the calculus-based course and even among students who have done the homework.

Firm grasp of the connection between the names and phrases on the one hand and the numerical calculations on the other begins to develop only after several homework and test exercises of the type illustrated in Section 5.16.

5.3 DESCRIBING EVERYDAY PHENOMENA

In addition to establishing the connection between numerical calculations and the names and phrases discussed in the preceding section, assimilation of the significance of the energy and momentum concepts is strongly enhanced by leading students to describe familiar phenomena, purely qualitatively, in the new vocabulary.

Many common experiences, confined to essentially rectilinear motions, lend themselves to such examination:

- (a) Throwing a ball vertically upward and catching it on its return (or allowing it to strike the ground).
- (b) Throwing a ball horizontally and having it caught by another person.
- (c) Throwing a ball horizontally so that it bounces back from a wall.
- (d) Throwing a ball of putty horizontally so that it sticks to a wall.
- (e) Jumping vertically upward and landing back on the ground.
- (f) Jumping vertically up and down on a trampoline.
- (g) Climbing a rope.
- (h) Walking and running (including starting, stopping, and maintaining uniform speed).
- (i) Pushing oneself off from a wall while standing on roller skates and coasting to a stop.
- (j) Approaching a wall on roller skates and stopping by cushioning the collision through flexing of one's outstretched arms.
- (k) Accelerating a car, maintaining uniform speed, and then slowing down and stopping.
- (l) Pushing a box along the floor in the presence of friction (including cases of speeding up, maintaining uniform velocity, and slowing down and stopping, and including the momentum and energy changes of one's own body as changes are imposed on the box).
- (m) Rowing a boat or driving it by means of engine and propeller.

- (n) Accelerating or slowing down a space vehicle by means of rocket propulsion in free space.
- (o) Stretching (horizontally) a spring fastened to a wall.
- (p) Stretching a spring horizontally by pulling its ends in opposite directions.
- (q) Sitting down on a soft chair.
- (r) Hanging an object on a spring fastened to the ceiling.

Students should be led to describe the energy and momentum changes taking place for all the interacting bodies, including the earth. It should be noted that some of the forces playing major roles in these phenomena are zero-work forces, and one must be very careful in the description of the energy transformations (cf. sections beginning with Section 5.6.)

Performance in such verbal descriptions is initially extremely poor among virtually all students, even among those of higher ability. The performance improves only with practice that is accompanied by evaluation and correction of the verbal statements. As the verbal performance improves, visible improvement in understanding is exhibited in problem solving and in grasp of the concepts themselves. Practice, as in other instances that have been discussed, must recur and must be spread out over time. Quick “remediation” is ineffective, especially with students below the highest ability levels; “mastery” is never attained on first encounter.

Fortunately, one can keep spiralling back in increasingly richer context if one simply remembers to take advantage of the opportunities for verbal description that arise in connection with circular motion, thermodynamics, electric and magnetic phenomena, waves and light, and with phenomena on molecular, atomic, and subatomic scales.

5.4 FORCE AND RATE OF CHANGE OF LINEAR MOMENTUM

Tempted by what seems (in hindsight) to be prescience on Newton’s part on the one hand and by the primacy of momentum in special relativity on the other, some textbook authors have been turning to emphasizing the second law form $\vec{F}_{\text{net}} = d\vec{p}/dt$, where $\vec{p} \equiv m\vec{v}$, in connection with development of the impulse and momentum concepts and in connection with conservation of momentum.

There is nothing wrong with this¹ except in those instances in which it is

¹What is being referred to here is the logical soundness of the concept. A serious question can be raised as to the pedagogical wisdom of adopting this approach too early in an introductory course in which the majority of students (even at the engineering-physics level) still find dv/dt mysterious and momentum bordering on the arcane.

implied that this form is more general than $\vec{F}_{\text{net}} = m\vec{a}$ because it allows for variation in mass as well as velocity.

The equation $\vec{F}_{\text{net}} = d\vec{p}/dt$ is valid, in general, only for *closed* systems. It happens to apply to one very special open system, namely, that in which the frame of reference is such that the velocity \vec{v} is not only the instantaneous velocity of the object undergoing change in momentum but is simultaneously the *relative* velocity of the incoming or outgoing material. It is *not* applicable to other open systems, especially, for example, the rocket.

These questions are now properly treated in many texts [e.g., Resnick, Halliday, and Krane (1992)], but incorrect or misleading statements are still sufficiently frequent to necessitate wariness and care on the part of the teacher.

5.5 HEAT AND TEMPERATURE

It is well known to most teachers that many students in introductory courses do not discriminate between the terms “heat” and “temperature” and tend to use the words synonymously when referring to thermal phenomena. This confusion will not go away by being ignored. It arises because, in the bulk of earlier student experience, the words have been used as though they were both primitives, with meaning obvious to everyone. The students have not been led to articulate simple operational definitions even though this can readily be done in elementary school science. Since this has not been taken care of earlier, it is necessary to give it some attention in introductory physics.

Fortunately, the task is not difficult because one can appeal to a large array of everyday experiences with thermal phenomena. A simple, basic strategy is to accept initial intuitive knowledge of thermometers and thermometer readings as primitives, not requiring elaborate definition to begin with.² One can then lead students to articulate descriptions of everyday experiences of which they are intuitively aware but which they have rarely made explicit:

- (a) When a bucket of hot water is put out in a room, the reading of the thermometer in the bucket always decreases, approaching the reading of the thermometer on the wall of the room.
- (b) When a bucket of cold water is put out in a room, the reading of the thermometer in the bucket always increases, approaching the reading of the thermometer on the wall of the room.
- (c) When two objects at different temperatures are brought in contact with each other, the temperature of the higher one always drops while that of the lower one rises until the two temperatures are equal.

²It can be pointed out that, ultimately, there is more to the concept of temperature than this but that, as with many other concepts, we go through successive steps of definition and redefinition in progressing to deeper insights from our initial, un-rigorous, starting point.

- (d) When two objects at the same temperature are brought together, no changes take place.

These descriptions, extended by others elicited from the students, can be used to generate the concept of “thermal interaction” between bodies initially at different temperatures and the concept of “thermal equilibrium.” Thus, one can cultivate explicit awareness of the prevalence of interaction and of the general trend toward thermal equilibrium in familiar phenomena.

The next step in the strategy is to lead students to recognize that, despite the powerful insight afforded by recognition of interaction and of the trend toward thermal equilibrium, the thermometer readings do not tell the whole story. Something else must be happening. Recognition of this fact can be elicited by drawing attention to the following observations and experiences in which thermometer readings are carefully controlled while other conditions are varied:

- (a) If we “insulate” the buckets of hot and cold water put out in the room by covering them with layers of cotton or of material such as that inserted in the walls of houses (or if we put the water in thermos bottles), we markedly alter the time it takes for the thermometer readings to reach equality with the reading on the wall. The initial and final thermometer readings are all still the same as they were without the insulation, but something has changed that the thermometers do not tell us about.
- (b) Similar time differences arise in the cooling down to outdoor temperature of two houses, with and without wall insulation, when heating is stopped inside. (It is surprising to find how many students are unaware of this and cannot really articulate the point and purpose of wall insulation.)
- (c) If we start with two containers of water, one large and one small, at the same temperature and place them to heat over identical burners, it takes a longer time and more fuel to elevate the temperature of the larger amount of water to the same final value as that of the smaller.
- (d) If we freeze or melt pure materials, the freezing or melting does not occur without interaction with a surrounding “bath” at either lower or higher temperature than that of the material undergoing the change. Despite the ongoing interaction during freezing or melting, however, the material undergoes no change in temperature at all until it is either all frozen or all melted (if students are not aware of this fact, it should be demonstrated).

All these illustrations, the last one especially dramatically, testify to the fact that thermometer readings do not tell the entire story of thermal interaction, that something else must be happening, and that an additional concept

(or concepts) must be invented.³ We give the name “transfer of heat from the higher to the lower temperature body” to the process of interaction implied by the preceding list of observations, and a recounting of this story constitutes a qualitative operational definition of the new term—which we all too soon, and all too cryptically, abbreviate to the one word “heat.”

Quantifying “transfer of heat” then follows in the usual way: Working with water as a convenient reference substance, one finds that (to first order of accuracy) equal masses of water, initially at different temperatures, end up at the mid temperature when mixed in thermal isolation from their surroundings. Unequal masses end up at a final temperature such that the ratio of the two temperature changes is the inverse ratio of the two masses. Thus, one verifies the possibility of using the masses and temperatures of quantities of water as a convenient way of assigning numbers to amounts of heat transferred. This leads to definition of the “calorie” and to the concept of “specific heat” of substances other than water. (The fact that the specific heat of water itself must be redefined since it varies with temperature, is a second-order discovery, illustrating the continual process of sharpening and refinement of concepts that takes place as the formation of a concept results in more sophisticated insight and then more precise quantitative measurement.)

At this juncture, it is important to return to cases such as that of the bucket of hot water cooling in the room. Few students explicitly recognize the room, or the house, as having a heat capacity, very much larger than that of the bucket of water but nevertheless finite. They must be led to recognize that the air in the room undergoes a temperature change very much smaller than that of the water but not zero.

It is worth emphasizing that we neither see nor measure heat transfer directly. (Few students notice this explicitly.) The quantities that we observe are masses and temperature changes, and we *infer* the amount of heat transferred from these observations. (Here is another opportunity to discriminate between observation and inference.)

For the sake of clarity and precision in forming and using the energy concepts in subsequent study, it is advisable never to speak of the “heat in a body,” even in the early stages of development of the concept. The term should be used only in connection with the process of transfer into or out of a body or system. This is the usage in thermodynamics, and it properly underpins the line integrals and inexact differentials associated with quantities of heat and work transferred.

As will be shown in Sect. 5.8, speaking of heat as though it resides in a body and implying it to be a function of state, raises severe impediments to clear formation of the concept of conservation of energy, even at an elementary level.

³As with the kinematical concepts discussed in Sections 2.5 and 2.9, here is still another opportunity to show students that scientific concepts are invented by acts of human imagination and intelligence and are not objects that are “discovered” as existing entities.

Furthermore, speaking of “converting work into heat” (when work is dissipated through frictional processes, for example) confuses the issue because heat is not actually transferred to the system in such a process. The work done on the system is converted directly into thermal internal energy *without* transfer of heat. The increase in thermal internal energy is found to be exactly equal to the heat transfer that *would* have been necessary to produce the same temperature change in the system. (These questions are discussed in more detail in Sects. 5.6 to 5.10.)

5.6 IMPULSE-MOMENTUM AND WORK-KINETIC ENERGY THEOREMS

Most textbooks start with

$$\vec{F}_{\text{net}} = m\vec{a} \quad (5.6.1)$$

and develop the impulse-momentum and work-kinetic energy relations, either as integrals of both sides of Eq. 5.6.1 with respect to clock reading t and position s respectively (in calculus-based courses) or in simplified, constant-force terms (in algebra-based courses). The initial development is, quite properly, carried out for the simplest possible case, namely, particles or point masses, and the equations take the familiar forms:

$$\int_{t_1}^{t_2} \vec{F}_{\text{net}} dt = m\vec{v}_2 - m\vec{v}_1 = \Delta(m\vec{v}) \quad (5.6.2)$$

or

$$\vec{F}_{\text{net}} \Delta t = \Delta(m\vec{v}) \quad (5.6.3)$$

and

$$\int_{s_1}^{s_2} \vec{F}_{\text{net}} \cdot d\vec{s} = \frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2 = \Delta\left(\frac{1}{2}mv^2\right) \quad (5.6.4)$$

or

$$\vec{F}_{\text{net}} \cdot \Delta\vec{s} = \Delta\left(\frac{1}{2}mv^2\right) \quad (5.6.5)$$

Both sets of results are perfectly valid and general for the situation for which they were derived, namely, the changes in motion of point masses. Serious difficulties begin to enter, however, when these relations are uncautiously extended to systems of interacting particles or to objects that are treated as continuous but are deformable or have other internal degrees of freedom. [Very common examples of such systems are: Motions imparted to our own bodies in walking, running, or jumping; propelling a car; discriminating between elastic

and inelastic collisions; and even pushing a (seemingly rigid) box over a floor with friction.]

Eqs. 5.6.2 to 5.6.5 are valid dynamical relations providing the velocities and displacements being referred to are those of the center of mass of the system under consideration. If one elects to attack extended systems, the significance of the center of mass and its role in the analysis should be developed at least plausibly if not with complete rigor. The general concept of conservation of linear momentum is properly associated with Eqs. 5.6.2 and 5.6.3.

Although Eqs. 5.6.4 and 5.6.5 are generally valid numerical relations (provided it is clearly maintained that velocities and displacements refer to the center of mass), serious conceptual and verbal problems begin to arise as soon as they are extended beyond application to point masses. The difficulties arise principally because these equations are presented as conservation statements for work and energy in analogy to the ways in which Eqs. 5.6.2 and 5.6.3 are conservation statements for linear impulse and momentum, the implication being that Eq. 5.6.4 is an initial version of the first law of thermodynamics and needs only refinement and extension to incorporate other forms of energy. Actually, Eqs. 5.6.4 and 5.6.5 are true work-energy statements for point masses but are not true energy statements for most applications to extended systems. Trouble arises because work and heat are forms of energy that are transferred across the boundaries of a system, and the system in question may be the single, deformable object with internal degrees of freedom.

That something is amiss begins to emerge as soon as we consider cases involving zero-work forces that impart acceleration to, and thus alter the kinetic energy of, extended bodies in various familiar situations:⁴ The force exerted on us by the ground when we jump vertically upward; the force exerted on us by the wall when we stand on roller skates and set ourselves in motion by pushing off from the wall with our hands; the horizontal frictional force exerted by the road on an accelerating car; and the frictional force exerted by the ground on us when we walk or run (in the absence of slipping) are all zero-work forces. Yet, in many texts, it is either said or implied that these forces do work and thus impart kinetic energy to the accelerating bodies, regardless of what might have been said earlier about forces that cannot be considered as doing work.

Another illustration that something is amiss stems from consideration of the very simple situation in which we push a box (taken as our system) at uniform velocity along a floor with friction. If we denote the force of sliding

⁴Virtually all textbooks and teachers properly point out, in connection with giving the name "work done by the force F " to quantities such as $F\Delta s$ or $\int \vec{F} \cdot d\vec{s}$, that zero work is done if there is zero displacement of the force in the direction in which the force acts. This is usually illustrated by appeal to situations such as our pushing on a rigid wall, where the displacement is zero, and carrying a suitcase at uniform velocity, where the displacement is orthogonal to the force. This aspect is, for the most part, competently embedded in instruction. It should be noted, however, that these illustrations are seldom, if ever, developed for situations in which kinetic energy is being imparted to a body.

uniform velocity along a floor with friction. If we denote the force of sliding friction by f and the displacement of the center of mass of the box by Δx_{CM} , Eq. 5.6.5 yields

$$f\Delta x_{\text{CM}} - f\Delta x_{\text{CM}} = 0 \quad (5.6.6)$$

implying that zero net work has been done on the box. Yet we know, from our more advanced knowledge of energy transformations, that a net amount of work has in fact been done on the box-floor system and that the equivalent of this amount of work has appeared as thermal internal energy, associated with the temperature rise exhibited by the box and the floor. In introductory courses we tend to paper over the apparent paradox by saying that the work done by us was “converted into heat,” with the implication—to the student—that heat has been transferred to the box. We know, however, that no heat transfer has taken place to increase the box-floor temperatures and that, whatever heat transfer does actually occur, is from the box and floor to the surrounding air.

The origin of these difficulties resides in the fact that Eqs. 5.6.4 and 5.6.5 are not really energy equations for anything but a point mass system without friction and without internal degrees of freedom. This does not mean that the equations are invalid or incorrect. They are numerically valid *dynamical* equations—connections among external forces (whether work-doing or not) and the displacement and velocity of the center of mass of the system to which $\vec{F}_{\text{net}} = m\vec{a}$ and its integrals are being applied. The contradictions being pointed to arise from a flawed and inconsistent invocation of the energy concepts.

Because Eqs. 5.6.4 and 5.6.5 are not true energy equations (except in a few very special cases), I shall, following Sherwood (1983), drop the “work-kinetic energy” terminology and refer to them as “center of mass” or CM equations. To emphasize this awareness, it might be more appropriate to write these equations in the form:

$$\int_{s_1}^{s_2} \vec{F}_{\text{net}} \cdot d\vec{s}_{\text{CM}} = \frac{1}{2}mv_{2\text{CM}}^2 - \frac{1}{2}mv_{1\text{CM}}^2 = \Delta \left(\frac{1}{2}mv_{\text{CM}}^2 \right) \quad (5.6.7)$$

or

$$\vec{F}_{\text{net}} \cdot \vec{\Delta s}_{\text{CM}} = \Delta \left(\frac{1}{2}mv_{\text{CM}}^2 \right) \quad (5.6.8)$$

making explicit the role of the center of mass coordinates and velocities and emphasizing the fact that the quantity on the left-hand side is *not* calculated from forces and their displacements around the *periphery* of the system (as must be done if the numbers so calculated are to obey a conservation law applicable to a system with internal degrees of freedom).

5.7 REAL WORK AND PSEUDOWORK

Since the concept of “energy” is not a primitive or intuitive one, students do not come to this aspect of physics with reasonable and deeply rooted preconceptions based on everyday experience—as they do with elementary dynamics. They come as *tabula rasa* and acquire certain misconceptions that are implanted in everyday speech and by many textbook presentations. These misconceptions then become hard to eradicate in later study.

The principal misconception planted in introductory physics is that the “work” quantity (force times center of mass displacement) appearing in the “work- kinetic energy theorem” (Eqs. 5.6.4 and 5.6.5), obtained by integration of Newton’s second law, is identical with the “work” appearing in the general law of conservation of energy, namely the first law of thermodynamics. (The reference to “real” work in the heading of this section is to the latter quantity rather than the former.) That this equating of the two “work” quantities is a misconception has been discussed in some detail in recent years [cf. Erlichson (1977); Penchina (1978); Sherwood (1983); Sherwood and Bernard (1984); Arons(1989)], but the necessary awareness has not yet penetrated many textbook presentations.

We must *discover* how to make the calculations that obey the conservation law, and we find that work done on or by a system (other than a single point mass) must be calculated by integrating forces and their accompanying displacements around the entire boundary of the system—the region to which the conservation law is being applied. Bridgman (1941) describes the operational problem in characteristic fashion:

Turn now to an examination of the W of the First Law. This W means the total mechanical work received by the region inside the boundary from the region outside [or delivered to the region outside from the region inside]. As in the case of [heat transfer] Q , this work is done across the boundary, and the evaluation of W demands the posting of sentries at all points of the boundary, and the summing of their contributions. In the simple cases usually considered in elementary discussions, the work received by the inside from the outside is of the simple sort typified by the motion of stretched cords or of simple linear piston rods. Our sentry can adequately report this sort of thing in terms of finite forces acting at points and finite displacements. In general, however, there will be contact of the material outside over finite regions of the boundary, and we become involved in the stresses and strains of elasticity theory. [Bridgman goes on to mention the “infelicities that result when we apply the notion of work to the sliding of two bodies on each other with friction.”]

Penchina (1978), joined by Sherwood (1983), suggests distinguishing between the two work quantities mentioned above by adopting the name “pseu-

dowork” for the quantity connected to displacement of the center of mass and reserving the name “work” for the quantity appearing in the first law of thermodynamics. Although this terminology has not yet become standardized, it is convenient because it does not completely sever the connection between the two quantities and because it does not resort to a radically new vocabulary. I shall therefore adopt it in the subsequent discussion. Those individuals who dislike this terminology are free to invent their own

5.8 THE LAW OF CONSERVATION OF ENERGY

The law of conservation of energy is, of course, not derivable from the dynamical laws of motion; it is an independent statement about order in the macroscopic world—one of the “principles of impotence” as the conservation laws and the second law of thermodynamics are sometimes called. The general conservation law, including heat transfer as well as kinetic, potential, and other energy changes, is a new statement that, in most cases, has very little to do with the CM equation.⁵ The way in which energy concepts are most frequently introduced in elementary physics courses does not make this fact sufficiently clear to the students.

Let us examine what aspects of the first law of thermodynamics might be comfortably incorporated into elementary physics in such a way that the treatment of energy concepts is correct and consistent but does not invoke mathematical and other complexities that are meaningless to students at that level.

The first law, stated in the familiar forms

$$dE = dQ - dW \quad (5.8.1)$$

or

$$\Delta E = Q - W \quad (5.8.2)$$

where E is called the “internal energy of the system,” is considerably more than just a statement of conservation of numbers calculated according to the “recipes” prescribed in the operational definitions of the various energy quantities. We have path-dependent quantities on the right-hand side (inexact differentials) and a new function of state (exact differential) on the left. In fact, the mathematical statement of the first law can be regarded as a definition in the sense that it turns out to be a law of nature that (1) there exists an energy quantity that is a function of state of the system (i. e., its changes are path independent) and that (2) we can calculate changes in this quantity in terms of path-dependent energy transfers that are not themselves functions of state. (It must be noted that up to this point we have defined only state

⁵See Section 5.6 for the definition of “CM equation.”

variables such as temperature, pressure, volume, mass, density, and so forth, without having any way of including an energy as a state variable.)

I do not propose trying to develop this sophisticated a perspective in elementary physics, but I do join Sherwood (1983) in urging that we adopt an essentially verbal translation that makes what we are actually doing in many commonly treated physical situations clearer to our students.

After we have shown that the dissipation of (real) work produces heating effects that are directly proportional to the amount of work dissipated (as Joule did in his famous experiments) and have suggested that future extensions will include electric, magnetic, and chemical effects in addition to the thermal and mechanical ones, we might say that we are asserting a new law of nature, namely that we shall always be able to find ways of calculating numbers (or “keeping the books”) such that the change in what we shall call the “internal energy of the system” is numerically equal to the sum of the quantities of heat and work transferred to it.⁶ In symbols, this would be translated as

$$\Delta E = Q + W \quad (5.8.3)$$

Note that this form alters the sign of W from that in the usual American convention in the first law.⁷ I suggest this convention for the sign of W , at least temporarily, since it would be very messy to give negative signs to work done by the force accelerating an object at a stage when students are having great difficulty with interpretation and use of algebraic signs in general. By the time one gets to thermodynamics with students who continue to that level, the convention can be changed without too much stress.

In preparation for future applications, let us list the internal energy changes with which we shall be concerned and specify the notation to represent each one:

- (a) Internal thermal energy change: ΔE_{therm}
- (b) Internal chemical energy change: ΔE_{chem}
- (c) Internal kinetic energy change: ΔE_{kin} . Internal kinetic energy changes would have subcategories such as:
 - (i) Translational: $\Delta E_{\text{kin, tr}}$
 - (ii) Rotational: $\Delta E_{\text{kin, rot}}$

⁶It is in this same spirit that Feynman introduces the energy concepts in the fourth of the Lectures on Physics [Feynman, Leighton, and Sands (1963)].

⁷In the most commonly used American convention, work is taken to be positive when it is done *by* rather than *on* the system. This is very likely motivated by the convenience of associating a positive value of dW with positive values of pdV for expanding fluid (or so-called chemical) systems. The convention is the reverse in much European and some American thermodynamic literature.

- (d) Internal potential energy change: ΔE_{pot} . Internal potential energy changes would have subcategories such as:
 - (i) Gravitational: $\Delta E_{\text{pot, grav}}$
 - (ii) Springlike in compression or extension: $\Delta E_{\text{pot, sp}}$
 - (iii) Electrical: $\Delta E_{\text{pot, el}}$
- (e) Miscellaneous internal energy changes: ΔE_{misc} encompassing emission or absorption of sound, radiation, or other messy interactions.

The general symbol ΔE in the first law (Eq. 5.8.3) would then stand for the algebraic sum of the various different internal energy changes specified in the above list:

$$\Delta E = \Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin}} + \Delta E_{\text{pot}} + \dots \quad (5.8.4)$$

To illustrate the implications of the approach being advocated, let us apply it to a very simple special case, namely the transfer of an amount of heat Q to a system in the absence of any doing of external work and in the absence of any internal energy change other than thermal. Equation 5.8.3, in combination with Eq. 5.8.4, would then give

$$\Delta E_{\text{therm}} = Q \quad (5.8.5)$$

This readily translates into the verbal statement that, under such circumstances, the amount of heat transferred to the system is equal to the change in thermal internal energy. Note that this statement uses the term “heat” only in connection with its transfer, as is systematically done in thermodynamics, and avoids the implication that heat “resides in,” or is a “property of” the system. (The subtlety that arises in connection with applying this terminology to everyday situations, such as those invoked in Section 5.5, is discussed in the next section.)

The first law, in the form of Eq. 5.8.3 or the preceding verbal statement, should be distinctly separated from the CM equation. It can be very neatly and simply applied to all the situations discussed in elementary physics, and, as I shall try to show in the following sections, such application greatly improves the clarity and consistency of the treatment of energy transformations in general, just as it does with the special thermal case illustrated above.

5.9 DIGRESSION CONCERNING ENTHALPY

Critical readers, especially chemists, will have noticed that the discussion, at the end of Section 5.8, of heat transfer and thermal internal energy change in the absence of doing of work lacks complete rigor. What is implied is a transfer of heat at constant volume of the receiving object or system. Although

such transfer is perfectly possible with gases, it is extremely difficult, if not actually impossible, with liquids and solids. Liquids and solids that expand on increase in temperature would have to be confined under enormous pressure to maintain constant volume, whereas water, which contracts in volume as its temperature increases anywhere between 0°C and 4°C, would have to be subjected to hydrostatic “tension.”

Virtually all the transfers of heat that we confront in everyday experience take place at constant pressure rather than constant volume, and some amount of work is inevitably exchanged with the surrounding atmosphere. Thus the amount of heat transferred is not strictly equal to the thermal internal energy change as defined by the first law. It was for this reason that the concept of “enthalpy” was invented: We define still another energy quantity, one that is a function of state and the changes of which are rigorously equal to the amount of heat transferred under constant pressure instead of constant volume.

This is achieved by applying a so-called Legendre transformation to the original expression for the first law:

$$dE = dQ - pdV \quad (5.9.1)$$

$$d(pV) = pdV + Vdp \quad (5.9.2)$$

If we add these two equations, we obtain

$$d(E + pV) = dQ + Vdp \quad (5.9.3)$$

The new function of state $E + pV$ is usually denoted by H and is called the “enthalpy of the system.” Equation 5.9.3 then becomes

$$dH = dQ + Vdp \quad (5.9.4)$$

and, rigorously speaking, the amount of heat transferred at constant pressure (as is the case in most everyday phenomena) is equal to the enthalpy change ΔH rather than the internal energy change ΔE . (The so-called “heat of chemical reaction,” dealt with in chemistry, is an *enthalpy change* between specified initial and final conditions, i.e., the same temperature and pressure.) The constant pressure heat capacity c_p is then defined as

$$c_p \equiv \frac{1}{M} \left(\frac{\partial H}{\partial T} \right)_p \quad (5.9.5)$$

where M denotes the mass of the system and T the temperature.

Under normal circumstances, with any system other than a gaseous one, the difference between ΔH and ΔE is extremely small and is legitimately neglected. With gases, one must be careful. [It is interesting to note that Julius Robert Mayer, who shares credit with Joule for discernment of the quantitative “equivalence” between work and heat, estimated the “mechanical equivalent”

not from thermal effects resulting from dissipation of work, as did Joule, but from comparison of the constant volume (c_v) and constant pressure (c_p) heat capacities of gases. He correctly interpreted the difference as reflecting the amount of work done on expansion as heat is transferred to the gas at constant pressure.]

In presenting this discussion of the enthalpy concept, I do so to try to clarify the subtleties involved and not to advocate its development in introductory physics courses. In the following sections, I shall use ΔE as an adequate approximation for ΔH . Formal introduction of enthalpy can perfectly well be left to the beginning of the more formal treatment of thermodynamics. (What we have here is still another illustration of how concepts are refined, redefined, and invented as knowledge and understanding deepen through successive approximations. It is worth noting that more than a century elapsed between Joseph Black's investigations of heat and the invention of enthalpy.)

A very few exceptional students occasionally get worried about the difference between constant volume and constant pressure processes. They can be directed to the more sophisticated view without confusing the issue for the others.

5.10 WORK AND HEAT IN THE PRESENCE OF SLIDING FRICTION

To illustrate the approach being suggested, consider the prototypical situation represented in Fig. 5.10.1: A block of mass m is accelerated from rest along a floor or table by an applied force F against a force of sliding friction f . The displacement of the center of mass of the block is denoted by Δx_{CM} and the final velocity by $v_{CM, f}$

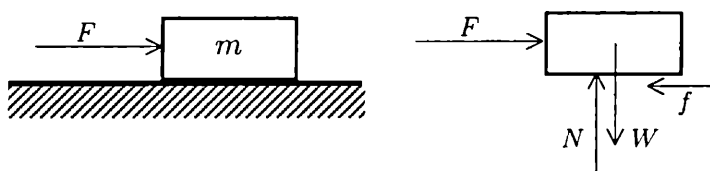


Figure 5.10.1 A block of mass m is accelerated from rest along a floor or table by an applied force F against the opposing force of sliding friction f .

We have two equations to apply to the changes taking place. One is the CM equation

$$F_{\text{net}} \Delta x_{CM} = \Delta \left(\frac{1}{2} m v_{CM}^2 \right) \quad (5.10.1)$$

and the other is the conservation of energy (COE) equation suggested in Section 5.8:

$$\Delta E = Q + W \quad (5.10.2)$$

For the situation in Fig. 5.10.1, the only relevant internal energy changes are thermal and kinetic translational, and Eq. 5.10.2 becomes

$$\Delta E_{\text{therm}} + \Delta E_{\text{kin, tr}} = Q + W \quad (5.10.3)$$

We can apply the CM equation 5.10.1 to the block itself, and we obtain

$$(F - f)\Delta x_{\text{CM}} = \frac{1}{2}mv_{\text{CM, f}}^2 \quad (5.10.4)$$

We must now carefully select the system to which we are to apply the COE equation 5.10.3. The block alone is not an appropriate system for this purpose because, although we can calculate the work done by the force F in displacing the center of mass, we cannot calculate the work done by the frictional force f at the interface. What happens at the interface is a very complicated mess: We have abrasion, bending of “asperities,” welding and unwelding of regions of “contact,” as well as shear stresses and strains in both the block and the floor. The work done on the block by the frictional force f at the interface is *not* simply $f\Delta x_{\text{CM}}$, and we are unable to deal fully with the quantity W on the right-hand side of Eq. 5.10.3.⁸

In this circumstance, we can take advantage of a possibility frequently available in application of the conservation laws: That of sweeping an area of ignorance under the rug by judicious choice of system. The system to choose in this instance is the *combination* of block and floor. (We could include the air also, but there is an insight to be gained by excluding it.)

If we apply the COE equation 5.10.3 to the system of block and floor, we no longer have to worry about the intractable situation at the interface since the frictional forces are now internal and no work is exchanged with the surroundings. Work is done on the system by the force F , and no heat is received from the surroundings. If there is any heat exchange with the surroundings (air) at all, it would be a *loss* following increase in temperature of the block and floor. Let us represent this by the positive quantity Q_{loss} . Substituting into Eq. 5.10.3, we obtain

$$\Delta E_{\text{therm}} + \frac{1}{2}mv_{\text{CM, f}}^2 = F\Delta x_{\text{CM}} - Q_{\text{loss}} \quad (5.10.5)$$

yielding

$$\Delta E_{\text{therm}} = F\Delta x_{\text{CM}} - \frac{1}{2}mv_{\text{CM, f}}^2 - Q_{\text{loss}} \quad (5.10.6)$$

⁸Sherwood and Bernard (1984) attack the problem of work done at the interface by adopting a plausible model and analyzing the consequences. They develop some interesting insights, and I recommend a reading of the paper. I would hesitate to take time for such an analysis, however, in introductory courses.

Interpretation of Eq. 5.10.6 yields the following insights: (1) The increase in thermal internal energy of the floor-block system is directly equal to that part of the work done by force F that does not go into the form of kinetic energy of the block, providing no heat is transferred to the surrounding air. (2) This work, which is said to be “dissipated,” is directly transformed into thermal internal energy; the increase in thermal internal energy does *not* result from a transfer of heat to the system. From this perspective, $F\Delta x_{\text{CM}}$ is real work done on the block-floor system, while $f\Delta x_{\text{CM}}$ is *pseudowork* done on the block. If F and f happen to be equal in magnitude and the block moves at uniform velocity, the pseudowork happens to be numerically equal to the real work, but that does not make the two quantities identical conceptually, and it does not make the net work done on the system equal to zero.

This approach is desirable because it helps one avoid misleading and confusing locutions about “converting work into heat” when no heat transfer is taking place. (The latter phraseology, inherited from the 19th century and never altered when other concepts were refined, helps implant, on the one hand, the misconception that heat resides *in* bodies, and, on the other hand, that heat is transferred to the system when work is dissipated.) One can now point out that the temperature change developed in the system by direct dissipation of work is equal to the temperature change that *would* have resulted from the transfer of an equivalent amount of heat even though no heat was actually transferred.

Given Eq. 5.10.6, one can explicitly idealize the situation by taking heat loss to the air to be zero, and one can apportion the total thermal energy increase between the block and the floor as may appear reasonable or as may yield upper or lower bounds on estimated temperature changes.

The reader might find it helpful to carry out the parallel analysis for the case, still dealing with the system in Fig. 5.10.1, in which the force F is removed, and the block coasts to a stop from an initial velocity $v_{\text{CM}, i}$. The CM equation becomes

$$-f\Delta x_{\text{CM}} = 0 - \frac{1}{2}mv_{\text{CM}, i}^2 \quad (5.10.7)$$

and the COE equation becomes

$$\Delta E_{\text{therm}} = \frac{1}{2}mv_{\text{CM}, i}^2 - Q_{\text{loss}} \quad (5.10.8)$$

No real work is done on or by the system, and all the internal kinetic energy is converted into internal thermal energy if there is no heat loss to the air.

Summary comment: The CM equation for a particular body gives an entirely correct numerical relation among dynamical quantities. Some of the terms in the equation are amounts of real work done on or by the body, but others are not and should be described as pseudowork. The proper interpretation of *energy transformations* comes from the COE and not from the CM

equation. The COE equation must be applied to a properly defined *system*. In some instances the pseudowork [a quantity that *looks like* an amount of work done (e.g., $f\Delta x_{\text{CM}}$), but is not a real work done by (or against) that force over the indicated displacement] is shown by the COE equation to be *numerically equal* to an amount of real work that was done by some other force (e.g., F) and was, say, dissipated. (Note that this kind of equality is analogous to another one that was discussed in Chapter 3: Although the downward force exerted on our hand by the body we are supporting is, in some circumstances, numerically equal to the weight of the body, the force on our hand is not the same force as the weight. The gravitational force exerted by the earth on the body and the contact force exerted by the body on our hand are two entirely different forces conceptually even when they are numerically equal.)

5.11 DEFORMABLE SYSTEM WITH ZERO-WORK FORCE

Consider the situation in which a person of mass m jumps vertically upward. The average normal force exerted by the ground on the person is denoted by \bar{N} . The center of mass of the person starts from rest and acquires a change in elevation Δh_{CM} and a velocity $v_{\text{CM},f}$ at the instant the feet leave the ground. For this situation the CM equation gives

$$(\bar{N} - mg)\Delta h_{\text{CM}} = \frac{1}{2}mv_{\text{CM},f}^2 \quad (5.11.1)$$

The quantity $\bar{N}\Delta h_{\text{CM}}$ is a pseudowork: It looks very much like an expression for work done on the person by the normal force \bar{N} , but it cannot be a real work done on the body since \bar{N} is a zero-work force; it undergoes *zero* displacement at the feet of the jumper. (If we could so easily derive energy from inert surroundings such as the ground, our nutritional requirements would be substantially less than they actually are.) The COE equation may show us that $\bar{N}\Delta h_{\text{CM}}$ is numerically equal to an amount of real work done by some other force, but that interpretation has to come from the COE and not from the CM equation. The quantity $mg\Delta h_{\text{CM}}$ can be interpreted, on the other hand, as an actual amount of work done by the jumper against the force of gravity if the jumper is taken as the system. It becomes an internal potential energy change in the jumper-earth system.

Let us first apply the COE equation to the jumper-earth system. No external work is done on the system, so $W = 0$. The only heat transfer might be a loss Q_{loss} from the jumper to the surrounding air. The relevant internal energy changes are: Chemical (the origin of the biological effects in the muscles), thermal (we warm up in such exercise), kinetic translational ($(1/2)mv_{\text{CM},f}^2$), kinetic rotational (flailing of arms), and gravitational potential ($+mg\Delta h_{\text{CM}}$). Hence,

$$\Delta E_{\text{chem}} + \Delta E_{\text{therm}} + \frac{1}{2}mv_{\text{CM},f}^2 + \Delta E_{\text{kin, rot}} + mg\Delta h_{\text{CM}} = -Q_{\text{loss}} \quad (5.11.2)$$

If we elect to ignore the flailing of the arms and the warming effects, this reduces to

$$\Delta E_{\text{chem}} = -\frac{1}{2}mv_{\text{CM},f}^2 - mg\Delta h_{\text{CM}} \quad (5.11.3)$$

Thus the source of the upward kinetic energy of the jumper and of the increase in the potential energy of the jumper-earth system resides in the decrease of chemical energy within the body of the jumper. This chemical energy is transformed through internal work-doing forces and displacements that cannot be described in quantitative detail. The CM equation 5.11.1 shows that the right-hand side of Eq. 5.11.3 happens to be numerically equal to $-\vec{N}\Delta h_{\text{CM}}$ despite the fact that \vec{N} is a zero-work force.

A useful insight is gained if we apply the COE equation to the system of the jumper alone rather than to the jumper-earth combination. In this case, external work $W = -mg\Delta h_{\text{CM}}$ is done by the gravitational force, and ΔE_{pot} must be taken as zero; otherwise this quantity would be entered twice. Ignoring thermal effects and internal rotation, the COE equation again gives Eq. 5.11.3

$$\Delta E_{\text{chem}} = -\frac{1}{2}mv_{\text{CM},f}^2 - mg\Delta h_{\text{CM}}$$

but the implications are different from what they were for the jumper-earth combination: The $mg\Delta h_{\text{CM}}$ term appeared, from the beginning, on the right-hand side of the equation as a work done by the gravitational force and did not appear on the left-hand side as an increase in potential energy of the system. This strongly reinforces the contention that we should only speak of potential energy changes of the interacting system (jumper-earth) and not speak of potential energy as residing in the elevated object (jumper) alone.

The reader is urged to set up the parallel analysis for the case in which the person stands on roller skates and pushes off from a rigid wall. The normal force \vec{N} exerted by the wall on the skater is again a zero-work force, and the source of the kinetic energy of the skater is the chemical internal energy change mediated by work done by unquantifiable internal muscular forces and displacements.

5.12 ROLLING DOWN AN INCLINED PLANE

The treatment of an object rolling down an inclined plane, with its usually sudden introduction of rotational kinetic energy and obscure condition of rolling without slipping, offers very great difficulty to many students. The usual consequence is memorization without understanding. The difficulties cannot be reduced to zero for the introductory level, but the physics can be significantly clarified by explicit separation of the CM and COE equations and careful interpretation of the content of the latter.

Consider the familiar situation in which a spherical or cylindrical object with mass m , radius R , and moment of inertia I rolls down an inclined plane (Fig. 5.12.1), starting from rest and acquiring center of mass translational velocity v_{CM} and angular velocity ω after center of mass linear displacement Δs_{CM} along the plane. The frictional force between the rolling object and the plane is denoted by f .

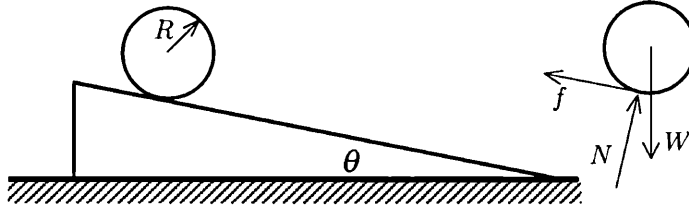


Figure 5.12.1 Ball or cylinder rolling down inclined plane. The plane is assumed to be undeformed in the region of contact with the rolling object.

Applying the CM equation yields

$$(mg \sin \theta - f) \Delta s_{\text{CM}} = \frac{1}{2} m v_{\text{CM}}^2 \quad (5.12.1)$$

Since W and Q are both zero, applying the COE equation to the object-plane-earth system gives

$$\Delta E_{\text{therm}} + \Delta E_{\text{kin, tr}} + \Delta E_{\text{kin, rot}} + \Delta E_{\text{pot}} = 0$$

which, on appropriate substitution, becomes

$$\Delta E_{\text{therm}} + \frac{1}{2} m v_{\text{CM}}^2 + \frac{1}{2} I \omega^2 - (mg \sin \theta) \Delta s_{\text{CM}} = 0 \quad (5.12.2)$$

Combining Eqs. 5.12.1 and 5.12.2 and solving for the thermal energy change gives

$$\Delta E_{\text{therm}} = f \Delta s_{\text{CM}} - \frac{1}{2} I \omega^2 \quad (5.12.3)$$

Thus the extent of thermal dissipation is determined by the comparative values of the two terms on the right-hand side of Eq. 5.12.3.

The additional information needed comes from the dynamical equation

$$\tau_{\text{net}} = I \alpha \quad (5.12.4)$$

where τ_{net} denotes the net torque acting around the axis through the center of mass of the rotating object, and α denotes the angular acceleration.

Integration of both sides of Eq. 5.12.4 with respect to angular displacement ϕ gives the rotational analog of the translational CM equation:

$$\tau_{\text{net}} \Delta \phi = \Delta \left(\frac{1}{2} I \omega^2 \right) \quad (5.12.5)$$

and specializing to the situation in Fig. 5.12.1 gives

$$fR\Delta\phi = \frac{1}{2}I\omega^2 \quad (5.12.6)$$

where $\Delta\phi$ denotes the total angle through which the object has turned since the rolling started.

Now let us examine what the energy equation 5.12.3 says about different conditions of rolling:

1 If the rolling takes place without slipping, the circumferential displacement (or unrolling) $R\Delta\phi$ must be equal to the linear displacement Δs_{CM} .⁹ Thus, for rolling without slipping, we have the condition

$$\Delta s_{\text{CM}} = R\Delta\phi \quad (5.12.7)$$

and Eq. 5.12.6 becomes

$$f\Delta s_{\text{CM}} = \frac{1}{2}I\omega^2 \quad (5.12.8)$$

Putting Eq. 5.12.8 into Eq. 5.12.3 gives

$$\Delta E_{\text{therm}} = 0 \quad (5.12.9)$$

showing that, for rolling without slipping, there is no thermal dissipation and that the decrease in gravitational potential energy of the system must be entirely reflected in the total kinetic energy of the rolling object. This is easily verified algebraically by going back to the earlier relations.

2 If we consider the case in which

$$\Delta s_{\text{CM}} > R\Delta\phi \quad (5.12.10)$$

the object must have been rolling along the plane without rotating as fully as it does in case 1 (i.e., there is some slipping). Equation 5.12.6 still holds, but it now indicates that

$$f\Delta s_{\text{CM}} > \frac{1}{2}I\omega^2 \quad (5.12.11)$$

Under these circumstances, Eq. 5.12.3 indicates that

$$\Delta E_{\text{therm}} > 0 \quad (5.12.12)$$

⁹Many students have a dreadful time seeing that rolling without slipping implies that $v_{\text{CM}} = R\omega$. Much of the trouble resides in their shaky understanding of the velocities. Students are helped if led to look at the problem concretely in terms of the total displacements, i.e., by comparing the linear displacement Δs_{CM} of the center of mass with unrolling of length of arc $R\Delta\phi$. Many students are unable to see the relation in the abstract and must be led to roll a cylindrical object themselves.

and some of the decrease in potential energy has been dissipated thermally. This is the nature of ordinary rolling with some slipping.

3 Now consider the case in which

$$\Delta s_{\text{CM}} < R\Delta\phi \quad (5.12.13)$$

(This means that the object is spinning *faster* than it would in nonslip rolling.) The pattern of the preceding analysis indicates that we will now obtain

$$\Delta E_{\text{therm}} < 0 \quad (5.12.12)$$

This is, of course, a result forbidden by the second law of thermodynamics. Such spinning (increase in kinetic energy at the expense of thermal energy) could occur only through input of work not included in the specification of the present system. One can use this example to illustrate the fact that we can easily visualize situations in which the first law is obeyed (i.e., energy is conserved), but the visualized change is nevertheless impossible because nature imposes some additional constraint. This eventually provides motivation for development of the second law.

5.13 INELASTIC COLLISION

Consider the case of a ball striking, and perhaps bouncing from, a rigid wall, neglecting effects in the vertical direction and assuming zero spin (Fig. 5.13.1). Let us apply the CM and COE equations to the interval between first contact between ball and wall and the instant at which the center of mass velocity of the ball is zero. With center of mass subscripts implied on displacements and velocities, the CM equation for the ball gives

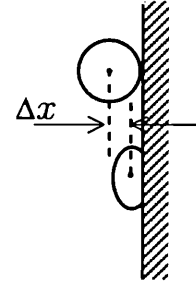


Figure 5.13.1

$$-\bar{N}\Delta x = 0 - \frac{1}{2}mv^2 \quad (5.13.10)$$

where \bar{N} denotes the average force exerted by the wall on the ball, Δx the displacement of the center of mass of the ball between contact and the instant of zero center of mass velocity, and v the center of mass velocity at instant of initial contact.

The $\bar{N}\Delta x$ term is a pseudowork, and we have no way of calculating the actual work done by the wall on the ball in the deformation accompanying the collision. The actual forces and displacements are very complex, as in the case of sliding friction, and cannot be described in such a way as to make calculation possible. The normal force \bar{N} is a zero-work force since displacements at the wall-ball interface are all parallel to the wall and are thus orthogonal to \bar{N} .

If we consider the ball-wall system, however, both W and Q are zero, and the general COE equation becomes

$$\Delta E_{\text{therm}} + \Delta E_{\text{kin, tr}} + \Delta E_{\text{pot, sp}} = 0 \quad (5.13.2)$$

where $\Delta E_{\text{pot, sp}}$ denotes any compressional potential energy that might be stored in the ball-wall deformations.

Rearranging Eq. 5.13.2, we obtain

$$\Delta E_{\text{therm}} = \frac{1}{2}mv^2 - \Delta E_{\text{pot, sp}} \quad (5.13.3)$$

Although Eq. 5.13.1 tells us virtually nothing, Eq. 5.13.3 is highly informative. It says that there is no increase in thermal internal energy, and therefore no thermal dissipation, if the increase in compressional potential energy in the system is equal to the initial kinetic energy of the ball. This defines the perfectly elastic collision (with coefficient of restitution of unity) since the center of mass speed (and kinetic energy) would be restored as the motion is reversed. (One could follow the reversal with the same equations we have written above but with appropriate changes in algebraic signs.)

At the other extreme, if the potential energy increase is zero (a ball of putty striking the wall), the increase in compressional potential energy is zero, and the increase in thermal internal energy of the system is equal to the initial kinetic energy of the ball (providing there has been no transfer of heat to or from the ball-wall system.) This defines the perfectly inelastic collision with coefficient of restitution of zero. In such instances, if one wishes to pursue the issue further (e.g., will the lead bullet with a given mass and initial velocity melt on striking the wall?), one must make explicit assumptions about the distribution of the thermal internal energy among the components of the system. The articulation of such assumptions becomes clearer and easier when one has explicitly defined the system and written an equation such as Eq. 5.13.3. The CM equation, on the other hand, is of no help at all.

Equation 5.13.3 also makes it possible to say something about partially elastic collisions, the cases intermediate to the two extremes examined above, if appropriate information is available or is assumed.

5.14 SOME ILLUMINATING EXERCISES

Because of the ubiquity of the experience, a very worthwhile exercise for the students is examination of real-work, pseudowork, and energy transformations associated with the propulsion of a car. If we accelerate a car of mass m from rest to a final velocity v , the CM equation gives, for the horizontal direction

$$(f_{\text{dr}} - f_{\text{res}})\Delta x = \frac{1}{2}mv^2 \quad (5.14.1)$$

where f_{dr} denotes the frictional force exerted by the road on the driving wheels, and f_{res} lumps together all the other forms of dynamical resistance to the motion. (Center of mass subscripts are implied on displacement and velocity.) Although f_{dr} is the accelerating force, it is a zero-work force (assuming no slipping), and the terms on the left-hand side of Eq. 5.14.1 are pseudowork, rather than real-work, terms.

Zero external work W is put into the car-road-air system, and there is zero transfer of heat Q to any other object. For this system, the COE equation gives

$$\Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin, tr}} + \Delta E_{\text{kin, internal}} + \Delta E_{\text{misc}} = 0 \quad (5.14.2)$$

One can make various idealizations and apportion various amounts to various terms. The main point to bring out, however, is that chemical internal energy, initially resident in the fuel, is transformed, through work done by internal forces, into kinetic energy of the car, thermal energy increases, and so on, and that no external work is done on the system despite the existence of an external accelerating force.

Summarizing the main thrust of our various examples: The CM equation is a perfectly valid *dynamical* statement, connecting the external force to the center of mass acceleration, but it is not necessarily a correct *energy* statement in the case of extended, deformable bodies or in the presence of friction. In the case of running, jumping, and accelerating a car, the external force accelerates the object on which it acts, but it does zero work on the system.

Sherwood (1983) suggests a number of other simple problems that are very illuminating conceptually and help point up the profound distinction between the CM and COE equations. The following problems are borrowed from his paper.

1 Sherwood ascribes the following to Michael Weissman: Consider the situation in Fig. 5.14.1 in which two frictionless pucks, connected by a string, are accelerated by the force F . (Note that this is a deformable system, that the force is displaced farther than the center of mass point, and that the pucks are assumed to undergo an inelastic collision on contact.)

The CM equation gives (for starting from rest)

$$F\Delta x_{\text{CM}} = \frac{1}{2}(2m)v^2 \quad (5.14.3)$$

Taking the system to be that of the two pucks, the external work done on the system is $F\Delta x$, and the COE equation gives (in the absence of heat transfer)

$$\Delta E_{\text{therm}} + \Delta E_{\text{kin, tr}} = F\Delta x \quad (5.14.4)$$

which, rearranged, yields

$$\Delta E_{\text{therm}} = F\Delta x - \frac{1}{2}(2m)v^2 \quad (5.14.5)$$

Note that the displacement Δx of the force F is greater than the displacement Δx_{CM} of the center of mass point and that the real work done on the system is greater than the pseudowork appearing in the CM equation. The difference between the two is dissipated in the inelastic effects and is equal to the increase in thermal internal energy. (A numerical problem could be posed by asking for the final velocity v and the amount of dissipated energy, given F , Δx , m , and the length of the string connecting the pucks.)

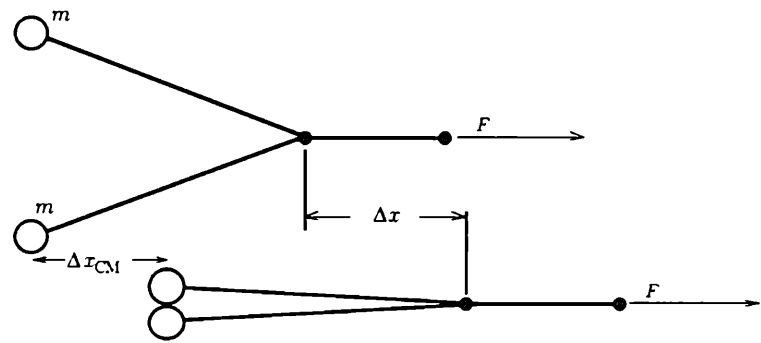


Figure 5.14.1 Top view of a system consisting of two frictionless pucks on an air table. The pucks are connected by strings as shown, and the external force F accelerates the system.

An alternative version of this problem would be that of starting with a bunched-up chain lying on a table. A horizontal force applied to one end of the chain stretches the chain out to full length and accelerates it along the table. Here again the displacement of the applied force is greater than the displacement of the center of mass of the chain, and thermal internal energy is increased through inelastic effects even in the absence of friction between the chain and the table.

2 Consider two identical blocks, each of mass m , initially at rest on a frictionless table and pulled away from each other by equal forces F (Fig. 5.14.2).



Figure 5.14.2 Top view of a system of two frictionless blocks on an air table. The blocks are accelerated by equal and opposite forces F .

If we take the system to be that of the two blocks, the CM equation gives

$$(F - F)\Delta x_{\text{CM}} = \frac{1}{2}(2m)v_{\text{CM}}^2 = 0 \tag{5.14.6}$$

Because of the zero on the left-hand side ($F - F$), this might seem to be a trivial result, but that is not so for the students. It emphasizes the fact that the center of mass of the system is not displaced and acquires no velocity despite the fact that work is done on the system and its individual parts acquire kinetic energy. In treating this situation, we normally apply the COE equation intuitively rather than formally, but a formal treatment in very simple cases does much to clarify the formalism for subsequent use in complicated cases where intuition fails us. In this instance formal application of the COE equation gives

$$\Delta E_{\text{kin, tr}} = W = 2F\Delta x \quad (5.14.7)$$

where Δx denotes the magnitude of the displacement of each block.

Substituting for $\Delta E_{\text{kin, tr}}$, equation 5.14.7 becomes

$$2F\Delta x = 2 \left(\frac{1}{2}mv^2 \right) \quad (5.14.8)$$

where v denotes the velocity acquired by each block. This is, of course, the result that we would write down intuitively, but it is actually justified only by the COE concept.

5.15 SPIRALLING BACK

An element referred to repeatedly throughout these discussions of physics teaching is the desirability of spiralling back so as to allow students to review or re-encounter important ideas and lines of reasoning in increasingly rich or sophisticated context. Following is an illustration of an opportunity for such spiralling back through use of momentum change in elastic collision in two dimensions.

The formula for centripetal acceleration of a particle moving at tangential velocity v in a circle of radius R

$$a_c = \frac{v^2}{R} \quad (5.15.1)$$

is usually derived initially in a kinematical treatment such as that in which one evaluates the acceleration associated with continual “falling” from the tangent line to the circle, or that in which one sketches the vector change in tangential velocity between two locations of the particle along the circular arc and then evaluates the rate of change of the velocity as the time interval tends to zero.

A simple dynamical derivation is given by Newton in Proposition IV of Book I of the *Principia*. Newton’s derivation has been accorded very little attention, possibly because it is given entirely in words citing proportional relations and is unaccompanied either by a geometrical figure or by algebraic equations. (In the Scholium following several corollaries to the proposition,

Newton acknowledges that “. . . by such propositions Mr. Huygens, in his excellent book, ‘De Horologio Oscillatorio,’ has compared the force of gravity with the centrifugal forces of revolving bodies.” Perhaps he did not play up the derivation because of Huygens’s priority.)

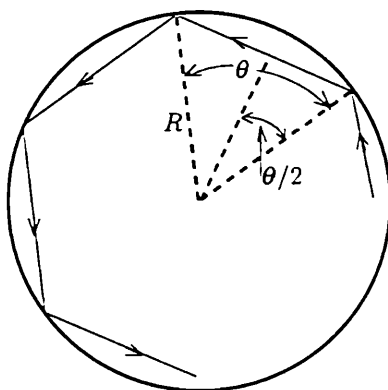


Figure 5.15.1 Particle of mass m moves at velocity v along sides of a polygon inscribed within a cylinder of radius R . The particle collides elastically with the walls of the cylinder at each vertex of the polygon. No gravitational effects are present.

Newton’s proof, translated algebraically and geometrically, runs as follows (see Fig. 5.15.1): Consider a particle of mass m moving at velocity v inside a rigid cylinder of radius R . Suppose at first that the particle moves along the sides of a polygon inscribed in the cylinder, undergoing successive elastic reflections from the wall of the cylinder at each vertex of the polygon. The momentum change Δp at each collision is given by

$$\Delta p = 2mv \sin \frac{\theta}{2} \quad (5.15.2)$$

The time Δt to traverse one side of the polygon is given by

$$\Delta t = \frac{2R}{v} \sin \frac{\theta}{2} \quad (5.15.3)$$

Since Δt is the time interval between successive impacts, the average rate of change of momentum is obtained by dividing Eq. 5.15.2 by Eq. 5.15.3, and, since the average force \bar{F} acting on the particle is equal to the average rate of change of momentum (second law)

$$\bar{F} = \frac{mv^2}{R} \quad (5.15.4)$$

(It is interesting to note that no small angle approximation is involved.) Newton thinks in terms of the collisions with the wall coming closer and closer together until, in the limit, the force exerted by the wall on the particle becomes continuous rather than consisting of a series of discrete impulses.

This analysis has very substantial pedagogical value—not as a *substitute* for the earlier kinematical derivation but as a powerful *supplement*. It gives the opportunity to come back to the centripetal acceleration and centripetal force ideas from an entirely new point of view after some time has elapsed

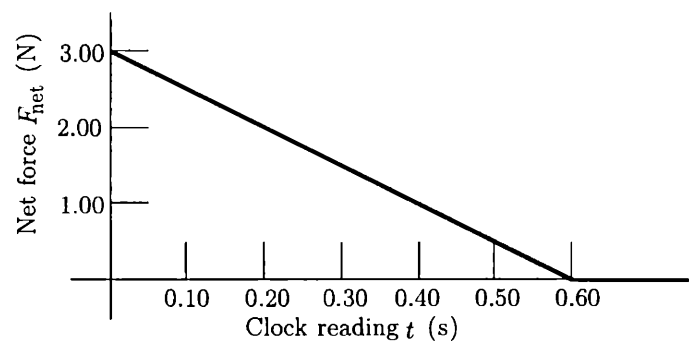
and after the momentum concept and its vector nature have been developed. The context is rich and conceptually significant: One makes use of the vector change of momentum in bouncing from the wall to obtain an important result and reinforce an earlier physical insight; the bouncing does not take place at normal incidence; the bouncing from the wall is not left as a seemingly sterile exercise in an end-of-chapter homework problem unconnected to other phenomena; a better foundation is laid for future use of such momentum change in formulating the kinetic theory of an ideal gas.

The order in which the different derivations of centripetal force or acceleration are developed is probably far less important than the fact of spiralling back. If the momentum concepts were developed prior to the discussion of circular motion, one could give the above derivation first and then reinforce it by coming back to the kinematical treatments later. It should also not be necessary to spend a great deal of valuable class time on the second treatment; the derivation could be assigned as a homework problem the solution of which is guided by a sequence of Socratic questions.

The Socratic questioning is needed by most students. At this stage, very few are ready to proceed with such an analysis without guidance. If the assignment is too cryptic, they will simply sit and stare at it without even putting pencil to paper. This disability is largely due to complete lack of opportunity to practice such thinking. Opportunities that are sufficiently simple analytically without being trivial conceptually are not very easy to find. That is what makes the opportunity outlined in this section especially valuable.

5.16 SAMPLE HOMEWORK AND TEST QUESTIONS

1 The net force (in newtons) acting on a glider on an air track (essentially frictionless system) varies with time as shown in the following diagram:



The glider has a mass of 0.850 kg. When the force is applied to the glider at clock reading $t = 0.00$, the glider has an initial instantaneous velocity of 0.150 m/s in the positive direction.

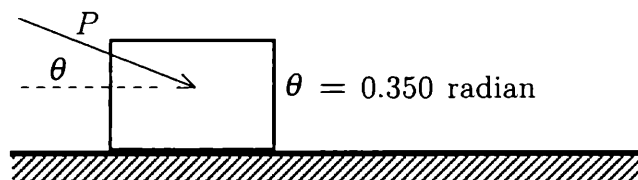
- (a) Describe, in words, what happens to the glider over the time interval between $t = 0.00$ and $t = 0.60$ s. (Use the vocabulary of impulse, momentum, work,

and kinetic energy rather than that of Newton's second law.)

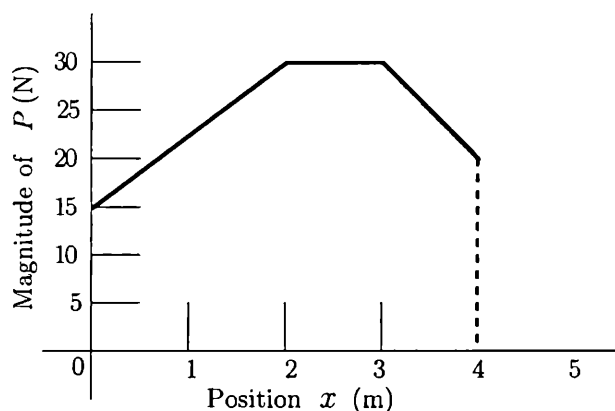
Perform the following calculations by making use of the concepts and relations referred to in your verbal description. In each calculation, describe your reasoning briefly.

- (b) Calculate the momentum of the glider at $t = 0.00$ s (i. e., the initial momentum).
- (c) Calculate the net impulse delivered to the glider by the applied net force between $t = 0.00$ and $t = 0.60$ s.
- (d) Calculate the change in momentum of the glider over this time interval.
- (e) Calculate the momentum of the glider at clock reading $t = 0.60$ s (i.e., the final momentum).
- (f) Calculate the instantaneous velocity of the glider at clock reading $t = 0.60$ s.
- (g) Calculate the initial and final values of the kinetic energy of the glider.
- (h) Calculate the change in kinetic energy of the glider.
- (i) Calculate the work that must have been done by the net force acting on the glider.
- (j) Calculate the potential energy change of the glider-earth system over the interval under consideration.

2 A block with mass 12.0 kg is being pushed along a horizontal floor by a force P as shown in the following diagram. The kinetic frictional force opposing the motion of the block is constant at a value of 15.0 N. At clock reading $t = 0.00$ s and position $x = 0.00$ m the block has an instantaneous velocity of 2.00 m/s.



The force P is applied to the block at the instant it is in position $x = 0.00$ m. The direction of P remains fixed, but its magnitude varies with position as shown in following graph:



Calculate the various quantities asked for in the following sequence. Use work, kinetic energy, momentum, and impulse arguments throughout; do not make use of Newton's second law. In each case, give a brief explanation of your calculation.

- Calculate the total work done by the force P in the displacement from position $x = 0.00$ to position $x = 4.00$ m.
- Calculate the work done against the frictional force in the same displacement.
- Calculate the work done by the net force acting on the block during this displacement.
- Calculate the *change* in kinetic energy of the block.
- Calculate the final kinetic energy of the block, that is, the total kinetic energy it possesses on reaching position $x = 4.00$ m.
- Calculate how far the block will slide beyond position $x = 4.00$ m if the force P abruptly drops to zero at this position and remains zero from there on.
- Calculate and describe any internal potential energy changes of the block-earth system that take place over the entire history of the motion dealt with in the preceding.
- Calculate and describe any internal thermal energy changes of the block-floor system over the entire history of the motion dealt with in the preceding.
- Calculate the velocity of the block at position $x = 4.00$ m.
- Calculate the momentum of the block at position $x = 4.00$ m.
- Calculate the momentum *change* imparted to the block over the interval between positions $x = 0.00$ and $x = 4.00$ m.
- Calculate the net impulse that must have been delivered to the block over this interval.

NOTE: The following suggests a lecture or class demonstration aimed at helping students establish concrete connections between observed motions and the symbolic statements embodied in the impulse-momentum and work-kinetic energy theorems.

[See Lawson and McDermott (1987) for a report of student difficulties with such connections and for a description of the apparatus used in their interviews.]

3 Gliders on an air track, pucks of different masses on an air table, or dry-ice pucks on a glass plate can be accelerated relatively slowly by means of the air stream emerging from the hose of the inverted vacuum cleaner system usually used to pump air through air tracks and air tables. (Low acceleration is desirable in order to make the sequence of events more clearly visible and apprehensible.) Attached to the hose opening are small strips of paper that, when blown out by the air stream, serve as spacers for maintaining a constant separation between the hose and the puck. This makes it possible to keep applying a fixed force to different objects.

- (a) Direct the air stream at a puck, started from rest, for a given interval of time, and repeat the demonstration once or twice to allow the students to acquire a clear visual impression of the time interval involved and of the velocity acquired by the puck. Then apply the air stream for shorter and longer intervals of time to the same puck and ask the students to compare the impulses delivered and the momentum changes imparted. (Emphasize that to “compare” means to indicate whether one quantity is greater, equal, or smaller than another—not to give numerical values.)
- (b) Direct the air stream for the same interval of time at two different pucks having clearly different masses and acquiring visibly different velocities. Ask the students to compare the impulses delivered and the momentum changes imparted.
- (c) Perform similar experiments in which the air stream is applied to the pucks over controlled displacements rather than over controlled intervals of time. Ask students to compare quantities of work done on the pucks and the changes in kinetic energy imparted.
- (d) Lawson and McDermott (1987) applied the same force to two pucks of very different mass over the space interval between two lines drawn on the table. They then asked the students (drawn from algebra-based and from calculus-based physics courses) to compare the momenta imparted to the two pucks and the kinetic energies imparted to the two pucks. The results obtained indicate the need students have for practice with the ideas afforded in this exercise.

Although it is useful to discuss at least some of the observations and interpretations as the demonstrations are being performed, it is also useful to leave some questions to be answered for homework or in group discussion. Many students need time to dwell on what they have seen and to think about the connections between the visible effects and the unfamiliar composite concepts (impulse, momentum, work, kinetic energy) to which they are being asked to accord physical meaning.

Chapter 6

Static Electricity

6.1 INTRODUCTION

For those of us who have, over years of experience, become intimately familiar with concepts and phenomena of electricity¹ and magnetism, it is all too easy to lose sight of how highly abstract this part of physics really is and how frustratingly difficult it turns out to be for many students. Its concreteness resides only at the level of observation of noncontact interactions that involve energy transfers through acceleration of objects, through deflections against opposing forces, or through thermal effects. We then construct abstract concepts and models that rationalize the observed effects. Because of the additional layers of concepts (such as “electric charge,” “like charges,” “unlike charges,” “electric current,” “potential difference,” “Lorentz force,” “field strength”) that are introduced, this conceptual structure is even further removed from the concrete manifestations than is the conceptual structure of mechanics.

It should not be surprising that students have very substantial difficulty with the most fundamental aspects of this subject: We are asking them to absorb an entirely new range of very sophisticated abstractions before they have fully mastered the related structure of mechanics. It is necessary to recognize this difficulty—to allow time for assimilation of new operational definitions, for confrontation with “How do we know . . . ? Why do we believe . . . ?” questions, and for establishing firm connections between the concepts and the phenomena. It is also important to keep invoking and applying the basic mechanical concepts (velocity, acceleration, force, mass, momentum, energy) at every opportunity.

Because most students have been hearing the language of electric and magnetic effects since childhood, many teachers and textbooks (especially at

¹I have placed the discussion of static electricity ahead of the discussion of current electricity, not because I consider this order essential, but because it is the order most commonly used in existing courses. One might perfectly well start with current electricity, as many teachers and some textbooks elect to do, and consider electrostatic phenomena afterwards. The same logical questions must be examined, however, in either case.

college level) tend to assume that the rudimentary ideas have been absorbed and understood. Unfortunately this is not the case. Many students have never directly observed electrostatic interactions. Many have never even seen the commonplace experiment in which a rubbed object attracts bits of paper, and many have never played with simple magnets. Many students, especially women, are afraid to touch small electric batteries because they have never had the chance to play with such objects and, at the same time, they have been conditioned to fear electricity.

Thus, many of the students coming to us at high school and college level have been hearing the vocabulary without ever having seen the phenomena that motivate it, and very few have examined the operational meaning of the terms with which they have become familiar. It is especially important to reexamine the basic vocabulary in any introductory physics course, whether at high school or college level. The failure to do this, the starting with the assumption that students must already “know” both the phenomena and the terminology, is responsible for a substantial portion of the subsequent difficulties students have with this subject matter.

6.2 DISTINGUISHING ELECTRIC, MAGNETIC, AND GRAVITATIONAL INTERACTIONS

Since discussion of electric, magnetic, and gravitational effects is usually introduced separately and at different times, whether it be in elementary school, high school, or college science, very few students have been afforded the opportunity to compare and contrast these phenomena operationally. The result is that many students in introductory physics courses, at any level, seriously confuse the terms and the effects. Interviews and test questions reveal that they do not distinguish clearly between electrostatic and magnetostatic effects: The terms are frequently used randomly and synonymously; many students will predict that north magnetic poles will repel positive electric charges; many expect other irrelevant and impossible effects; some students have the impression that gravitation is some sort of electric or magnetic phenomenon. Very few students have ever had direct, hands-on experience with the phenomena under circumstances in which they are considering the different manifestations at the same time and can compare the similarities and differences explicitly.

This aspect of learning and understanding is mentioned at this point not in order to advocate such discussion in class prior to the experiences outlined in the following sections, but in order to alert the teacher to the importance of raising the question and returning to it periodically as further operational distinctions can be added to the list. The point is to keep the students explicitly conscious of the differences as they observe electric and magnetic phenomena and compare them with gravitational ones.

It should be kept in mind that the only familiar aspects of gravitation are

the weight and the free fall of objects. The minute gravitational interactions of ordinary bodies and gravitational interaction on the cosmic scale both involve abstract extension of the concept of “gravity” to situations beyond the realm of direct experience. Keeping the various phenomena in mind simultaneously enriches the context and enhances understanding of the vocabulary and remembering of the effects.

6.3 FRICTIONAL ELECTRICITY, ELECTRICAL INTERACTION, AND ELECTRICAL CHARGE

Never having personally observed the phenomena and interactions that lead to formation of the concept of “electrical charge,” but having heard the words used in and out of school since early childhood, many students use the term without knowing what it means and what it does not mean. To many students, charge is some invisible kind of substance that may be smeared on things or that may drip out of household sockets.²

In an introductory physics course, students should be led to retrace at least some of the observations that led to the formation of the concept of “electrical interaction” resulting from frictional effects, then to formation of the concept of “charge,” recognition of the mobility of charge, and finally to realization that there are no more than two varieties of charge (or two “charge states”) and to the operational definitions of the terms “like” and “unlike.” Teachers who are not familiar with the historical sequence as it developed in the 18th century will find the story very instructive and illuminating even though they might not want to follow all of its intricacy in the class work. [Especially useful references to this end are Whittaker (1951) and Roller and Roller (1941). An abbreviated version is given by Arons (1965)].

Those students who have never observed the attraction of bits of paper or fibers to a charged plastic rod or comb should be led to do so. In making the observations, they should be led to think about gravity at the same time. They should articulate the fact that the interaction is different from the gravitational since gravity has nothing to do with the rubbing of objects and persists unchanged by the rubbing effects. They should also perceive that the electrical interaction is enormously stronger than the gravitational one since the bits of paper are given a high upward acceleration despite the downward pull of the entire earth. Such differences are sufficient to justify a new name, and the term “electrical interaction” can be introduced.

The term “charge” is introduced as the name for the property acquired by the interacting objects—a property that seems to “leak away,” can be restored by rubbing, is transferable from one object to another by contact, is

²Readers aware of the humorous writings of James Thurber may note that these views are very similar to those ascribed by Thurber to his mother in *My Life and Hard Times*. The ascription is not idiosyncratic.

highly mobile on metallic objects, and so on. It must be explicitly emphasized that “charge” is an abstract construct and not a substance—that it is the name of a property that we infer from observed interactions and that we never come to know what charge “is” or how it “works” any more than we know what gravity “is” or how *it* “works.” (Many students, having heard about electrons in connection with electricity and electrical charge, and having no idea whatsoever as to what the term “electron” means or where it comes from, labor under the misconception that, by knowing about electrons, they must know the nature of electrical charge. They must eventually be disabused of this misconception. One approach that makes a dramatic impression on most students is to lead them to the perception that the only reason we know electrons to be negatively charged is, in the final analysis, that we observe electron beams to be repelled by rubber rods that have been rubbed with cat fur.)

The terms “conductor” and “nonconductor” (or “insulator”) should be given explicit operational definition in connection with those demonstrations and experiments that deal with the mobility and transferability of charge. This is more or less competently done in most presentations, and it is not necessary to dwell on it here.

6.4 ELECTROSTATIC EXPERIMENTS AT HOME

Since very few students have observed electrostatic effects at first hand, there is much to be gained by encouraging them to perform relevant experiments themselves in addition to passively watching any lecture demonstrations that may be performed. (This is not meant to diminish the importance of the lecture demonstrations; these are very worthwhile and should certainly be performed, but they do not obviate the need for firsthand involvement.)

Fortunately, it is possible to perform effective and useful electrostatic investigations at home, and such homework assignments (as opposed to routine pencil-and-paper work) prove engaging to many students, especially if they are given adequate guidance.

The key to such electrostatic experiments is ordinary white (“MagicTM”) Scotch^R tape. If a strip of tape two or three inches long is stuck down on virtually any smooth, dry surface and is then quickly pulled off, it turns out to be charged—strongly with some materials and relatively weakly with others. (One end of the strip should, of course, be folded over so as to provide a “handle” that does not stick to the surface and allows the strip to be handled without sticking to the fingers.)

If one such strip of tape is stuck to the back of another and the doubled tape is stuck to a smooth surface and pulled off, the two pieces of tape turn out to be oppositely charged when they are pulled apart.

One end of such a charged strip of tape can be stuck to some readily available support (e.g., the edge of a table, the bottom of a lamp shade, the slat

of a chair back, etc.), and the suspended strip then becomes an electroscope leaf that can be observed to react to the presence of other charged objects. If the two oppositely charged strips of tape are suspended not far from each other, one has two reference strips capable of interacting with, and checking, the charge on other objects.

In a lecture demonstration, students can be shown how to set up such electroscopes and how to go about conducting an investigation of the interaction of charged objects. The first steps of investigation might involve charging other strips of tape by similar procedures and observing their interactions with the electroscope strips, but it should be made clear that the investigation should not stop with strips of tape. Many other household items (plates, glasses, containers of all kinds, toothbrush handles, combs) can be charged by rubbing with various materials (silk, wool, fur, cloths made of synthetic fibers, plastic bags or wrapping films), and the rubbed objects can be tested for interaction with the electroscope strips. One possible test object is one's own body after scuffing over a rug in dry weather, the test being made by bringing one's finger near each of the oppositely charged electroscope tapes.

A report of the French Academy in 1733 points out that "when an electrified body is brought close to the face [or arm], it causes a sensation like that of encountering a cobweb." This is a sensation students should be led to experience and interpret.

Some of the systematic investigations that might be conducted and questions that students should be led to address are outlined in the following sections.

6.5 LIKE AND UNLIKE CHARGES

Very few students, even those who may have acquired some substantially correct knowledge of electricity in prior schooling, can give an adequate operational definition of what is meant by "like" charges. If asked for the meaning of the term "like," virtually all of them say "like charges repel" and regard that as a definition. Very few textbooks pay any attention to this nontrivial linguistic problem. Many textbooks confound the issue by assuming that everyone "knows" that there are two kinds of charge and that like charges repel while unlike charges attract. They fail to comprehend the necessity of defining the technical terms "like" and "unlike" and of asking why we believe there to be more than one variety (or state) of electrical charge and why we also believe that there are not more than two. Only a very few textbooks handle these questions clearly and correctly.

Furthermore, a few students seem to acquire the (not very explicitly articulated) idea that there is really only one variety of charge. When questioned closely, they indicate a belief that repulsion between two bodies reveals that both are charged while other objects, attracted by either of these bodies, are actually without charge. This is, of course, a subtle issue since charged

bodies do always attract uncharged ones, and we eventually rationalize this phenomenon by inventing the concept of “polarization.” Even though most students do not reveal this misconception openly (they have memorized the dictum that there are two varieties so efficiently that the difficulty does not seem to arise), very few students are able to outline experimental evidence supporting the view that there are at least two “charge states” rather than one.

These elementary, but conceptually very significant, questions can be resolved by raising them explicitly while guiding the students through home experiments such as those suggested in Section 6.4. Guidance is best provided through a sequence of Socratic questions tailored to the level of preparation and sophistication of the group. If the students are simply turned loose to “investigate” the electrostatic interactions on their own, very little happens except among an extremely small number of exceptional individuals. The investigation has to be guided with sequences of questions that lead the students to make relevant observations and draw inferences. The questions must be carefully structured, however, so as to lead the students from one insight to another without giving the whole story away by simply asserting the end results.

The first step involves leading students to recognize that two strips of tape pulled off the same surface always repel each other and then to the extension that identical objects, charged in an identical way by being rubbed with the same material, always repel each other. This provides a first cut at the operational definition of “like.” (An extension or redefinition of the concept into the completely general assertion that “like charges repel” comes later when we begin to recognize that there seem to be only two varieties of charge.) The recognition of two distinct charge states (or varieties of charge) comes, of course, from the observations that (1) rubber rods that have been rubbed with cat fur always repel each other; (2) that glass rods rubbed with silk always repel each other; and (3) that the glass and rubber rods then always attract each other. Finally one notes (4) that if either set of like rods were uncharged, they would not interact.

One way of approaching these insights in the home experiments is to set up the two electroscope strips of tape that are obtained when first stuck together back to front and then pulled apart after having been pulled off some smooth surface. (As indicated in Section 6.4, the two strips are found to attract each other strongly.) Students can be led through the logical sequence outlined in the preceding paragraph by use of a second set of strips prepared in the same manner as the first two.

Students, in their investigation of subsequent effects, can then bring up other charged objects to each strip and check whether the interaction is attractive or repulsive. Very few students will spontaneously discern and articulate the key, common element in the resulting observations. They must be led into articulating the insight that, if a charged object attracts one of the two strips,

it always repels the other one, and vice versa. They must then be led to recognize the importance of what is *not* observed but can be imagined: Namely that another object that has been rubbed with a different material is never found to repel both of the electroscope strips that attract each other. Neither do we observe a charged object that attracts both of the strips, but here one must be very careful to be sure that the test object is indeed charged, since any large uncharged object does attract both strips. (See further discussion below.)

Because no one has ever found a charged object that either attracts or repels both of the two strips of tape that attract each other, we come to believe that there are only two “charge states.” It is not possible to “prove” that there are only two any more than we can “prove” the law of conservation of energy. We accept the assertion of a regularity only because a violation has never been observed.

Now that we are limited to two “states” of electrical charge, we can extend the meaning of “like” to embrace all repulsive interactions and invoke the terms “unlike” or “opposite” to cover all attractions (with the exception of the subtlety outlined below). The names “positive” and “negative” can then be logically introduced in any one of the usual ways.³

As they make their observations, students should simultaneously be led to pay attention to the qualitative strengths of the interactions they observe. They should become explicitly aware of the fact that they intuitively assess the strength of the interaction by the force exerted on one of the interacting objects (e.g., the deflection of the strip of tape). They should recognize that (in the light of Newton’s third law) there must be an equal and opposite force acting on the other object, but that the effect is usually too small to be observed directly as an acceleration or deflection. (The deflection of both strips of tape is readily observed, however, if a hand-held charged strip is brought near one suspended at the edge of the table.) They should be led to

³From a carefully logical point of view, it is worth noting that there is a subtle distinction between recognizing the existence of two distinct “charge states” and asserting the existence of two distinct “varieties” of electrical charge. As far as macroscopic phenomena are concerned, we can account for the observed phenomena in either of two ways: (1) We can visualize the displacement of a single, conserved, imponderable fluid (as Franklin did) leading to one charge state in which the fluid is present in excess of normal amount and another state in which there is a deficiency, or (2) we can visualize two distinct varieties or kinds of charge, as we have become used to doing.

It is wise to remember that, on the macroscopic scale, we cannot really distinguish between the two models and that either one is thus equally valid. The choice of the “two distinct varieties” model is eventually forced on us by *microscopic* rather than macroscopic phenomena. How much of this detail one should belabor with students at introductory level is something the individual teacher must decide, given the existing constraints of time and coverage. One should not, however, suppress those students who try to pursue the “one fluid” model initially. They are not wrong, and they are in good intellectual company, given the history of the concept. They should be led to see how the modern “two varieties” picture came to be accepted rather than the “excess” and “deficiency” picture.

recognize that the force increases markedly with decreasing spacing between the interacting objects. They should recognize that, when each of two charged objects is separately brought near a charged strip of tape, the force exerted on the tape at the same distance of separation is frequently quite different, stronger for one object and weaker for the other. From this they should begin to draw the implication that one might eventually measure the charge state of an object by the force exerted on another “standard” charged object.

In the sequence of electrostatic experiments with the oppositely charged Scotch^R tape electroscopes, there is a very important subtlety that students must be led to recognize. Any relatively large object that has not been rubbed with something else (and hence is presumably uncharged) will attract both charged strips of tape. One’s own hand if brought close, for example, will attract both of the electroscope strips. Students must become aware of the fact that, although the repulsive interaction is always clear cut, an attractive interaction must be carefully checked and cross-checked to determine whether it is really an interaction between unlike charges or whether it is the ever-present attraction between a charged object and an uncharged one.

In the home experiments, there is a good way of showing that, if an object acquires one variety of charge on being rubbed, the rubbing material acquires the opposite charge. Normally (as with the hand-held fur and silk of the conventional lecture demonstrations), charge leaks off the rubbing material so rapidly that it is almost impossible to show that it becomes charged. If one puts one’s hand, however, into a small plastic sandwich bag and charges some object by rubbing, the plastic bag retains its charge far more effectively than do the silk and fur. When tested against the electroscope strips, the bag will show a charge opposite to that of the rubbed object. One must, however, take the hand out of the bag before approaching the electroscope; otherwise the attraction of the large uncharged object (the hand) will overwhelm the interaction one seeks to observe.

There is ample room in these apparently simple home experiments for careful and skillful experimental work and for keen observation. As one performs the investigations oneself and devises ways of guiding students of different levels of sophistication and preparation through a fruitful sequence, one begins to see very vividly how subtle the unravelling must have been for the 18th century investigators and why it took them so long to recognize explicitly that there were two distinguishable charge states and not just one.

6.6 POSITIVE AND NEGATIVE CHARGES; NORTH AND SOUTH MAGNETIC POLES

Many students labor under the misconception that the names “positive” and “negative” for the two varieties of electrical charge are in some way necessary or inevitable, or that they literally refer to something in excess and to something

missing. If the teacher desires, it is worth explaining how the names came into use historically through Benjamin Franklin's plausible (but, from the modern point of view, mistaken) model [see references such as Cohen (1941) and Roller and Roller (1957) for the details.] It should be made clear to the students in any case, however, that the names are perfectly arbitrary and could just as well have been chosen to be "red" and "blue" or "George" and "James" or "charming" and "revolting" (to parody some of the tongue-in-cheek choices that have been made in recent years in particle physics.) The rubber rod, rubbed with cat fur, can be established as the primary classical reference for negative charge in lecture demonstration and can be used to calibrate the students' Scotch^R tape electroscopes for home use.

Once the positive-negative terminology has been established in this way, it is well worth making a brief digression to the behavior of ordinary magnets. Some students have played with magnets in their earlier years, but it is surprising how many have not. Furthermore, even those who have played with magnets have rarely done so under circumstances in which they made systematic and guided observations rather than engaging in random manipulations that were never ordered or organized conceptually. Many students, for example, have played with magnets only to the extent of lifting tacks, nails, or paper clips with a single magnet but have never explored the interaction between two magnets and are astonished when they first observe repulsion. It is also a fact that some students may have seen such effects demonstrated under various circumstances, but the experience has been entirely vicarious, and they have never set up the interactions themselves or felt the forces with their own muscles. It is important for such students to feel for themselves the attraction between two magnets on being pulled apart in appropriate orientation or the repulsion when pushed toward each other.

Here again, students can be guided into systematic home-based experiments after a few illustrations in lecture and under the guidance of notes that supply a suitable sequence of Socratic questions. Students can be encouraged to purchase their own magnets for home experiments or suitable kits can be issued from the classroom.

The term "magnetic pole" can be developed operationally from observations of the differences between interactions associated with the ends of bar magnets and their middles. It should be made clear, preferably by direct demonstration, that magnetic poles cannot be separated by breaking the magnet and that all simple magnets, however large or small, are always bipolar. The question can then be posed as to whether there might be a valid role for the terms "like" and "unlike" as there was with electrical charge. The best reference frame for developing the operational definition is the historical reference frame, namely the earth itself.

Many students are aware of the interaction between magnets and a compass needle, but initially it is best to stay away from the compass because the compass needle is itself only an ordinary magnet. Many students do not

realize this, and others all too easily lose sight of the fact that the primary reference in magnetic nomenclature is the earth rather than the compass.⁴

The students should be led first to establish the north-south direction wherever they are located. They can then take two rod or bar magnets (first having checked that they are actually both magnets and that one of them is not simply an unmagnetized object) and suspend them on strings, with a suitable yoke to maintain a horizontal orientation. The students should observe for themselves that each of the magnets, after coming to rest, is approximately lined up with the north-south direction. They should then mark the “north seeking” poles in some way and check the interactions between all of the combinations of magnet ends. From such observations they can be led to give an operational meaning to “like” and “unlike” poles and to articulate the observation that like poles always repel each other while unlike poles attract. They can also see the origin of the names “north” and “south” for the respective north- and south-seeking ends.

Following such a sequence, most students will readily conclude that the compass is simply a pivoted magnetized needle and will recognize that the earth itself must be a huge magnet. Most teachers are aware of the confusion students exhibit in connection with nomenclature concerning the poles of the earth, but this confusion stems principally from the fact that students have never had the opportunity to go through a sequence of observation, reasoning, and invention of names such as that outlined above. They have been constrained to passive reading or listening and have only tried to memorize the vocabulary without sharing the experiences on which it is based. When they make the observations themselves, they can see through the tricks of the vocabulary and can recognize that the so-called north magnetic pole of the earth must be the south magnetic pole of the earth-magnet.

Playing with magnets provides an opportunity to extend the list of operational distinctions among electrostatic, magnetic, and gravitational forces. Students should be led to try the effects of magnets on their Scotch^R tape electroscopes and to ascertain that there is no interaction other than the previously noted attraction observed in the presence of any sizable object. This paves the way for eliminating misconceptions such as repulsion between a north magnetic pole and a positive electric charge, and so on.

Students should be led to make detailed lists of both the similarities and the differences among the effects, for example, there are two kinds of magnetic poles and two kinds of electrical charge, but there is no known bipolarity in

⁴Teachers should be alert to the fact that many students do not know how the directions “north” and “south” are established relative to the surface of the earth (see Section 1.14), and that many labor under the misconception that the compass defines these directions. It must be made clear that the directions are defined by astronomical phenomena and that the compass turns out to be a convenient secondary device only because it was discovered that the needle tends to align itself approximately in the astronomically defined north-south direction.

gravity; magnetism, as it is being explored with common permanent magnets, is confined to certain metals and has nothing to do with frictional effects as does electrical charge; magnetism is not “drained away” from an iron magnet when it handled (as charge is immediately drained away when one touches any isolated charged conductor); gravitational interaction is ever present, in all materials, without having anything to do with either frictional effects or magnetization; and so on. Even fairly knowledgeable students have never thought of making such a list and, when the question about specific operational distinctions is raised, have trouble giving illustrations of observed behavior and articulating the essential differences.

6.7 POLARIZATION

At about this juncture, it is helpful to direct students' attention to the fact that accumulating experience with electrical phenomena points to a very deep role for electrical charge in the structure of matter, even though experiments up to this point do not reveal just what that role is: (1) If we accept the notion that the architecture of matter is discrete (atoms and molecules) rather than continuous, the apparent presence of both varieties of charge in all materials hints that attraction between positive and negative might be what holds the discrete entities together; (2) sparks jumping through air (and, in the process, discharging charged bodies) hint at some sort of breakdown of structures ordinarily neutral and containing both kinds of charge—the breakdown resulting in conduction not present initially; (3) conduction, charging by frictional contact, and charging by induction all testify to different degrees of mobility of electrical charge within various materials.

Given awareness of these implications of the observations being accumulated, one can fruitfully go back to the observation that there is always an attraction between charged and uncharged objects. On the basis of what we now know about the ubiquity and mobility of electrical charge, can we provide a plausible description or model of how this interaction might arise?

The plausible model is, of course, provided by visualizing displacement of charge within the neutral body when in proximity to the charged one: Opposite charge within the neutral object is then closer to the charged body than is the like charge, and the net result is attraction between the two objects. We bring to bear here almost everything we have learned so far about electrostatic phenomena, and we give the name “polarization” to the displacement of charge that is induced in the neutral object.

This well-known set of ideas is summarized in such detail at this point, not because it is in some way obscure or difficult, but because of the unfortunate manner in which it is developed in many existing textbooks. A very common approach is to assert that attraction between charged and uncharged bodies arises “because” of the phenomenon of polarization, and this is usually done without first reviewing and summarizing the existing knowledge as is done in

the first paragraph of this section. Few students are prepared, at this stage, to unscramble the logical sequence on their own. Furthermore, they take “because” statements very literally: When the scientist says that such and such happens “*because* of so and so,” many students take such statements as an absolute accounting of “why” the phenomenon occurs. The scientist has come forth with an a priori *reason* that the student feels he or she could not possibly have conceived, and the revelation is obediently memorized without comprehension of whence it came.

I have, on various occasions, characterized such a presentation as “backwards science”; it forgoes the opportunity to give students a more realistic view of how science works, and it obscures the intelligibility and motivation of the model.

The appropriate starting point is not the model but the observed *fact*, and the observed fact is that of attractive interaction between charged and uncharged bodies. We seek to account for this interaction in terms of what we now know, and the concept of “polarization” is deliberately invented to provide a plausible account. The concept is invented a posteriori and not a priori. This is the “forward” rather than the “backward” sense of the development.

This context can be enriched by leading students to consider the analogous effect in the case of permanent magnets: The lifting by a magnet of a string of tacks or paper clips and the general occurrence of attraction between a magnet and any unmagnetized piece of iron. The logical sequence is essentially the same as that in the case of electrical polarization, but students should also be led to note the differences as well as the similarities in the observable phenomena.

Without some direct help, many students find the *process* by which electrical polarization occurs very difficult to visualize and comprehend. They should be assisted in visualizing possible processes in both conductors and nonconductors. Since charge is visualized to be nonmobile in nonconductors, the appropriate model is that of inducing electric dipoles. (Many texts introduce this idea but usually so cryptically that students are mystified and tend to memorize without comprehension.) Since charge is visualized to be mobile in conducting bodies, polarization is to be visualized in terms of actual displacement or separation of charge.

In the latter case, it is important to lead students to the realization that the model accounts for attraction by a charged body regardless of whether one imagines positive charge to be mobile, or negative charge to be mobile, or both varieties to be mobile simultaneously. Many students are puzzled by this and are very reluctant to accept the indeterminacy. They expect to have the “right answer” immediately, and they also find it hard to accept the idea that one cannot obtain a unique answer from the experiments and observations. (In the same way, students, when they begin to do some naked eye astronomy and observe the behavior of the sun, moon, and stars, expect to “see” that the earth and planets revolve around the sun and that the earth rotates on

its axis. They find it disconcerting to confront the fact that the heliocentric and geocentric models account for the observations equally well and that it is impossible to distinguish between them at this level of observation.)

Furthermore, the electrical situation is heavily confounded by the fact that many of the students have heard about “electrons” (without understanding anything about the origin and meaning of the term), and they have been misled into explaining virtually all electrical effects in terms of motion of “electrons” whether or not the concept is relevant. Students should have the opportunity to see that macroscopic observations and experiments never do resolve the question of what variety of charge is displaced in conductors and that, for the time being, any one of the three possible models is equally valid. Also, students who labor under the fixed delusion that all conduction is by electrons should be apprised of the fact that, in ionized gases and in electrolytes, there exist both negative ions (that are not electrons) and positive ions, and that conduction in such systems takes place through migration of both kinds of ions, and sometimes by positive ions alone.

6.8 CHARGING BY INDUCTION

Many students in introductory physics courses find charging by induction to be a difficult and mystifying process and have great trouble visualizing it correctly and making correct predictions. Most of this difficulty stems from inadequate grasp of the concept of polarization discussed in the preceding section. If the idea of polarization is clearly motivated and developed, much of the difficulty with charging by induction goes away (providing it is made clear that the entire system, object plus other body in contact, becomes polarized in the presence of the charged object causing the induction.)

Exercises with charging by induction should, however, require that students visualize the effects in terms of the mobility of either kind of charge, or even in terms of mobility of both. Such practice is readily cultivated if it is made clear that test questions might require use of any one of the models chosen at random.

In analyzing situations involving charging by induction, students should be required to draw pictures of charge distributions at various successive stages in the sequence of operations and to describe, in their own words, what is happening. It is under the requirement of describing each step in words that understanding is most effectively generated.

6.9 COULOMB'S LAW AND THE QUANTIFICATION OF ELECTRICAL CHARGE

Given the fact that electrical charge is not a ponderable substance and, up to this point in the sequence of development, is simply the name for a certain state

in which objects attract or repel each other, it is far from obvious to students that the “quantity” of electrical charge is measurable. Eighteenth century investigators (prior to Coulomb) had no way of measuring the quantity of charge, although they qualitatively recognized more and less highly charged states by the intensity of the observed effects, such as sparks and shock, and they compared charge states roughly by the number of turns that had been given the electrostatic generator prior to having the generator make contact with the Leyden jar or other body accepting charge.

It was Coulomb who first showed convincingly that charge might be quantified by measurement of the force of interaction. In the famous torsion balance experiments in which he established the inverse square law, he also examined the interaction between charged conducting spheres at fixed separation as the charge state was altered in a systematic way: Having set up a situation in which the charged spheres on the torsion balance were, say, repelling each other, he brought an identical uncharged conducting sphere in contact with one of the two interacting spheres and observed that the force between the charged spheres was now half of what it was initially. On bringing the uncharged sphere in contact with the second of the charged spheres, he observed that the force of interaction was now one quarter of what it had been initially.

Arguing from symmetry (i.e., that identical spheres should share charge equally if charge is indeed a systematically measurable quantity), one supposes that the quantity of charge on each of the interacting spheres was successively cut in half in the above procedure. Since the force of interaction was also cut in half at each step, it is plausible to infer a direct proportionality between observable force and the product of the two interacting quantities of charge. Thus charge becomes truly quantified not by qualitative observation of intensity of sparks and shocks or by number of turns of an electrostatic machine but by means of a previously established numerical scale, namely that of force. This leads directly to the usual statement of Coulomb’s law induced from these observations: That the force of interaction between point charges is directly proportional to the product of the two quantities of charge (and inversely proportional to the square of the separation). In the old c.g.s. system of units, each of the interacting spheres was defined as carrying one unit of charge when the force between them was one dyne at a separation of one centimeter between centers.

Thus, in electrostatics, prior to the connection to current electricity and its chemical effects, support for the notion that quantity of electrical charge was measurable came from the halving of the force in Coulomb’s experiments. The *PSSC* film “Coulomb’s Law” (narrated by Eric Rogers), presents an excellent demonstration of these effects on large scale, showing that the force between charged spheres varies inversely as the square of their separation and also showing that the force at fixed separation is cut in half each time one of the spheres is brought in contact with an identical uncharged sphere. These effects can also be demonstrated quantitatively and very effectively by setting

up a Coulomb's law experiment on the platform of an overhead projector. (The warmth supplied by the projector helps maintain dryness that minimizes leakage of charge during the course of the demonstration.) A transparent graph paper on the platform supplies a scale. The system consists of one pithball on a bifilar suspension so that the ball can swing away from equilibrium position, the displacement from equilibrium position being a measure of the force acting on it. A second pithball is fastened to the top of a short insulating rod on a small stand of transparent plastic that can be moved along the projector platform. The stand can be moved around, bringing its pithball to within different distances of the ball on the bifilar suspension. The two balls are set up so as to be quite close to the platform of the projector, and their images, as well as that of the graph paper, appear on the screen. The inverse square effect can be nicely demonstrated up to the point at which small separation of the spheres leads to appreciable distortion of the charge distributions by induction. The halving of the force can be shown by touching one of the charged spheres with an identical uncharged one.

In developing this story for students, it is important to emphasize the fact that Coulomb's law is not "derived" mathematically or "proved" physically, any more than $F = ma$ is derived or "proved." Postulating the law on the basis of the very limited observations described above is a matter of *inductive* and not deductive reasoning. Ultimate acceptance of the law and belief in its validity reside in the fact that it has been tested over a vast variety of situations and applications and has always been found to "work" and never found to fail. Furthermore, the quantification of charge via Coulomb's law turns out to be consistent with quantification through definition of electric current and through observation of chemical effects (i.e., deposition of material in electrolysis). In this context it is also illuminating, for example, to show students why it is that careful investigators have taken great pains and invested substantial effort in testing the accuracy of the inverse square relation. Williams, Faller, and Hill (working at Wesleyan University in 1971) showed that, if the exponent deviates from exactly 2, it does so by less than 2×10^{-16} .

Finally, this is an opportune point at which to remind students that, just as Coulomb's law is induced rather than derived or proved, so also are the Newtonian laws of motion, the law of gravitation, and the conservation laws. When one is traversing these developments for the first time, it is very easy to lose sight of subtle logical aspects of this variety; yet the awareness is crucial to understanding the nature and limitations of scientific knowledge, as well as the essential difference between science and mathematics.

It is also worth noting that Coulomb, in the infancy of the science, talks not about electrical "charge" but about the electrical "masses" of his charged spheres. It is obvious that "electrical mass" has nothing whatsoever to do with "inertial mass." By analogy, one can dramatize the idea and help students discriminate operationally between inertial and gravitational mass by temporarily referring to the latter as "gravitational charge." (See Section 3.9)

6.10 ELECTROSTATIC INTERACTION AND NEWTON’S THIRD LAW

The encounter with electrostatic and magnetostatic forces provides a valuable opportunity for spiralling back of the variety advocated repeatedly in these notes, namely a re-encounter with Newton’s third law after some time has elapsed since its use in mechanics. Those teachers who are inclined to believe that the issue has been settled through clear exposition, and through the numerous exercises assigned in the first encounter, would do well to try the following problem on students who have begun the study of electrostatics:

The following diagram [Fig. 6.10.1] shows uniformly charged spheres firmly fastened to, and electrically insulated from, frictionless pucks that ride on an air table. One sphere carries a charge of $+2.0 \times 10^{-8}$ C while the other sphere carries a charge of $+6.0 \times 10^{-8}$ C. The pucks, with the spheres they are carrying, are free to accelerate along the table. Draw separate free-body force diagrams of the spheres and of the pucks, showing all the forces acting on each of the four objects. When forces are equal, show the arrows as equal in length. When forces are not equal, show the larger forces with longer arrows to some reasonable scale.

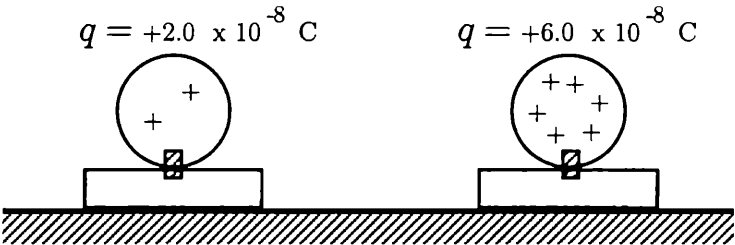


Figure 6.10.1 Charged conducting spheres are pinned to nonconducting frictionless pucks and are allowed to accelerate freely on an air table.

When I gave this problem on a diagnostic pretest to a group of 40 second-year students who were beginning an introduction to modern physics after having completed a full year of introductory calculus-based physics, 65% of the group showed the electrical force acting on one sphere to be three times as great as the electrical force acting on the other, and 85% showed no horizontal force of interaction between the spheres and the pucks (60% also had something wrong with at least some of the vertical forces). I submit that this experience dramatically illustrates the importance of giving students repeated exposure to qualitative questions of this variety. It is only after several encounters, in altered context, spread out over time, that the concepts register with the large number of students who do not operate at the front edge.

In addition to recognizing that the forces are still equal and opposite even when the charges have different magnitudes, the students should also recognize

that, when the charged rod they are holding attracts bits of paper or repels a charged pithball, forces are being exerted on the rod even though they cannot be sensed directly. This fact is commonly overlooked, and the situation is not connected with the one in which forces between, and accelerations of, small interacting objects are clearly revealed.

Still another question involving Newton's third law can be fruitfully raised at this juncture: Suppose we have two charged spheres attracting or repelling each other; in accordance with the third law, the forces are equal and opposite. Suppose we now suddenly displace one of the spheres. Does the force exerted on the other sphere change instantaneously (i.e., does electrostatic interaction operate as an "action at a distance" in the manner Newton assumed was the case with gravity)? If the change is not instantaneous, and an interval of time elapses before the change, it must be that the third law does not hold during that interval. What are the implications of this line of thought? (See Section 3.13 for comments on the initiation of this discussion.) It should be noted that "closure" need not be provided immediately. Many students have been conditioned into expecting an immediate pat answer for any question raised, however subtle and sophisticated it might be and, in the case of the present question, the answer was long in coming. Faraday spent the last years of his life vainly trying to observe such time delays, and the question came to be resolved only with Maxwell's invention of field theory and the recognition that the law of conservation of momentum has primacy over Newton's third law.

6.11 SHARING CHARGE BETWEEN TWO SPHERES

The sharing of charge between identical conducting spheres was made use of in the discussion of Coulomb's law and the quantification of charge (Section 6.9). Distribution of charge between spheres of different radii is usually discussed in connection with the concept of capacitance. In all these instances, contact is made between the outside surfaces of the spheres, and it is shown that the ratio of the charges acquired by the spheres is equal to the ratio of the radii.

After it has been shown that there is no charge residing on the inside of a hollow conductor, and students should have this idea firmly registered, it is illuminating to give them a question such as the following: Suppose we charge a large hollow sphere and then insert a small uncharged sphere into the space within the large one through an opening in the latter. Now we connect the small sphere to the interior of the large one with a wire. What will be the final charge distribution between the two spheres?

It has been my experience that many students blindly go back to the relation for the charge distribution between the two spheres making external contact and apply the same formula to the small sphere making contact with the inside of the large one. They fail to invoke what has been learned about the

interior of the hollow conductor, and they do this even after vivid demonstrations of the “ice pail” experiment, especially if a bit of time has elapsed since the discussion regarding zero electrical field in the interior of a hollow conductor. They show better retention of the latter concept after having made the mistake indicated above. This is another illustration of the value of spiralling back in a later context.

6.12 CONSERVATION OF CHARGE

If one seeks to cultivate and enhance “critical thinking” among the students, an important avenue for doing so resides in leading them to address the “How do we know . . . ? Why do we believe . . . ? What is the evidence for . . . ?” questions that arise at the most fundamental levels of concept formation and of perception of order in natural phenomena. The questions raised in the preceding sections underpin genuine understanding of, and critical thinking about, electrical phenomena. Adoption of the view that electrical charge is conserved is an integral part of the structure. What is the evidence for this principle? The insight that electrical charge is conserved is far from obvious in the performance of macroscopic electrostatic experiments, and it should not be surprising that this realization came rather late in the historical sequence.

In 1747, Benjamin Franklin conducted a series of simple, ingenious experiments with Leyden jars. Two identical jars were charged in the same way by simultaneously touching the “hooks” (contacts to the internal foil) to the glass tube of an electrostatic generator while the external foil was grounded. Franklin observed that if the outer coatings were connected to each other (by holding the jars with his fingers, his body being the conductor), nothing happened when the hooks were brought in contact. However, when he held one jar by the hook and the other by the outside coating, and then brought the coating of the first in contact with the hook of the second, a spark occurred, and the jars were completely discharged.

This experiment, together with variations of it, led Franklin to surmise that the “electrical fluid”⁵ was merely shifted from one body to the other in frictional contact and that the total quantity was always conserved. It is true that the experiments are crude and qualitative, but Franklin’s intuition was sound: the conjecture influenced other investigators and proved to be very fruitful in subsequent research.

The Leyden jar demonstrations are easily done in class or lecture, and questions regarding their implications make excellent homework assignments—assignments that give students a chance to consider the interpretation of real

⁵Franklin held a one-fluid model of the charging process and (fallaciously) convinced himself that glass acquired the excess of the fluid while the material with which it was rubbed acquired the deficiency. From this heuristic model came the terms “positive” and “negative.”

physical phenomena, with all the attendant caveats and uncertainties, in addition to the conventional numerical problems with precanned parameters.

6.13 ELECTRICAL FIELD STRENGTH

The concept of “electrical field strength \vec{E} at a point in space” as defined by

$$\vec{E} \equiv \frac{\vec{F}}{q_{\text{test}}} \quad (6.13.1)$$

is well known to offer difficulty to many students. Most of the difficulty appears to reside in the definitional aspects, but, since there are several of these, the effect is compounded.

First, the use by most textbooks of the ordinary equals sign (=) instead of the identity (or definition) symbol (\equiv) deflects attention from the fact that the idea being invented resides on the right-hand side of Eq. 6.13.1 while the left-hand side is simply a name for this idea. (See discussion of different kinds of “equalities” in Section 3.23.) This needs to be emphasized several times before it sinks in, and the identity symbol helps in maintaining the emphasis.

Next, we have the mystique of the ratio. This is still a matter of discomfort to many students, and they fail to ask themselves what the ratio means, that is, what is the verbal interpretation? They hope to be able to manipulate the formula without facing the verbal issue. It is necessary to lead them to articulate the fact that we are talking about the force per unit positive charge and not about force alone. Obvious though this may seem, it is frequently lost sight of. Since it helps to have done this kind of thinking more than once and to have started in some more familiar situation, quite a few textbooks now mention the parallel concept of “gravitational field strength,” that is, gravitational force per unit mass, and lead the student to recognize that this is just another way of talking about the familiar concept denoted by g . This parallel illustration is well worth invoking.

Next, there is the question of why we want to invent such a definition in the first place. The point, of course, is that we wish to create a number that is an intrinsic measure of the force field to be described and that must therefore be made independent of numerical properties of our detecting device. Although this objective is explicitly stated in many presentations, it is frequently put forth so briefly and casually that most students miss it. Students should be led into participating in the invention of the concept and should have to articulate the purpose and the interpretation in their own words. It is through such articulation that they begin to perceive that a concept is being created for a specific purpose and cease to regard Eq. 6.13.1 as a “formula” received through revelation.

Next, students should explicitly summarize and describe the operations underlying Eq. 6.13.1, that is, they should be able to describe step by step

what they themselves would *do* to obtain the number denoted by \vec{E} . It is in such a description that students are constrained to indicate that a force is to be measured at a point in space and that an electrical force cannot be observed without placing a charged particle at that point. This description becomes the best place to emphasize that the test charge must, in principle, be so small as not to disturb or rearrange the charges creating the field under consideration, and that \vec{E} therefore denotes force per unit charge without actually being measured as a force on one whole coulomb.

One might hope that, having been led through the preceding sequence, all students would understand the meaning of “electrical field strength” and would pursue the solving of conventional numerical problems without a hitch. The trouble, however, is that, when starting in on a problem (especially one involving the calculation of a field strength involving superposition of effects from two or more point charges), only a very few students take themselves back through the meaning of \vec{E} on their own initiative. The majority sit staring at the problem and wishing to discern how to substitute in the formula when what they should be doing is asking themselves “What does \vec{F} stand for? What does q_{test} stand for? What would I do to obtain the respective numbers?” Only then do they proceed, for example, to find the vector sum of the forces exerted on the test charge by other point charges in the vicinity. It is necessary to help students form the habit of stopping and asking themselves such questions whenever they find themselves in difficulty with definitional statements appearing in symbolic form.

6.14 SUPERPOSITION

“Each point charge in a distribution of charges contributes to the total value of \vec{E} at a given point as though it (the point charge) alone were present.” This seems like a simple and reasonable statement, and it is—except for the fact that many students are confused by a few normally unspecified aspects.

If one is concerned with a single idealized charge distribution in empty space (e.g., a set of point charges, a line charge, a sheet of charge), the only difficulty that arises is associated with the idea of setting up an integration for the resultant effect of a continuous distribution. Here the difficulty usually resides in the fact that many students do not yet clearly associate integration with the process of addition needed in these circumstances. They have been *told* about limits of sums, but they have rarely, if ever, had to use such language themselves in describing what they are doing in specific integrations. They have executed algorithms of integration without talking about or interpreting them. Many do not yet recognize the evaluation of an area as an addition process and, even if they do, they do not carry the addition concept over to the process of finding the total field strength. They must be led to describe what is happening in their own words.

Significant conceptual difficulties with the superposition concept arise when the effects of more than one simple array of charge are being evaluated. First, students must be helped to absorb the subtle idea that the insertion of a set of charges in a region where mobile charges (especially those on conductors) are already present will, in general, lead to a rearrangement of the charge distributions and that the superposition principle applies only to the final rearranged state. (I am not concerned here with dielectric phenomena or other sophisticated aspects of polarization but only with simple situations of the kind that arise in an introductory course.) For example, if one is dealing with the field produced by two or more charged conducting spheres and is treating them as point charges concentrated at their centers, students should be articulately aware that the latter assumption holds only to the degree that the interactions among the spheres are not sufficiently strong to distort the initially spherical charge distributions residing on their surfaces. They should also be aware that the true resultant field is that due to whatever final static distribution is actually attained.

A second difficulty with superposition is that many students are puzzled by what happens when the situations under consideration imply the presence of physical objects rather than just disembodied charge distributions. In considering the fields produced by the charged spheres mentioned in the preceding paragraph, for example, they find it very hard to believe that they can treat the system as though the metal spheres themselves were absent. They believe that “one sphere would not let the field from the other sphere ‘pass’ through it.” Similarly, after it is shown that the field on either side of an infinite plane sheet of charge is uniform to infinity, they are unwilling to accept the argument that superposition of the fields of the two plates of a capacitor gives an approximately zero field outside the plates. The unwillingness arises, not from the idealizations and approximations involved, but rather from the robustly held idea that the field of one plate could not possibly “penetrate” through the other plate.

Many students hold these misconceptions on an a priori basis without ever having heard of electrostatic shielding. The misconception is reinforced, however, in those students who *have* heard of shielding. They tend to associate the shielding effect with inability of an external field to penetrate the shield rather than with the cancellations due to the redistribution of charge on the shield itself.

Genuine understanding of the superposition principle can be developed only through recognition and careful discussion of these conceptual difficulties. Students must be helped to realize that the principle is arrived at by inductive reasoning, that it is accepted because it “works,” that it is a fact of nature that, once the charge distribution has attained equilibrium, intervening spheres and capacitor plates play no role, and that electric field contributions cannot be thought of as “blocked by” or “passed through” material objects.

Chapter 7

Current Electricity

7.1 INTRODUCTION

Research has been showing that the most basic concepts underpinning simple direct current (d.c.) circuits offer very serious difficulties to many students and that certain misconceptions are widely prevalent [Arons (1982); Cohen, Eylon, and Ganiel (1983); Fredette and Clement (1981); McDermott and Shaffer (1992)]. As in the case of static electricity, the learning problems are aggravated by the remoteness of the underlying phenomena from direct sense perception. The observable effects are not easily linked to abstractions such as “electrical charge,” “current,” and “energy.” Since students are aware that batteries “run down” and that one “uses” household electricity, they believe that “something is used up” in electric circuits, and, to many of them, the most reasonable thing to be “used up” is “electricity” itself. Furthermore, the concept of “potential difference” is difficult enough in its application to electrostatic fields; its relevance to, and applicability in, electric circuits is even more obscure for most students.

The structure of the concepts underlying Ohm’s law, and the operational definitions entailed, are just about as intricate as those underlying Newton’s second law,¹ but there has been far less epistemological discussion of the electrical case. Textbooks show a very wide range of differences in order and mode of presentation. Research on student learning difficulties is still in its infancy. At the present time, no one mode or sequence of presentation emerges as pedagogically superior to any other. Teachers are on firm ground if they choose any logically sound mode which is most congenial and with which they feel most secure.

Regardless of the logical structure of presentation, however, conceptual difficulties arising in the treatment of current electricity tend to be glossed over

¹I do not mean to imply that there is a close analogy between Ohm’s law and the second law of motion, although some authors attempt to exploit such an analogy. In logical structure and range of validity, Ohm’s law is much more like Hooke’s law for ideal springs than it is like Newton’s second law.

much too rapidly and superficially in introductory courses at all levels. As a result, very few students (even in engineering-physics courses) develop understanding of the phenomenology of simple circuits. They are usually tested on numerical manipulations involving Ohm's law, while the ability to solve the conventional problems has only very weak correlation with understanding of the physical phenomena taking place [Arons (1982); McDermott and Shaffer (1992); Shaffer and McDermott (1992)]. This situation can be remedied only by leading the students to think more carefully and more frequently about the qualitative aspects of the phenomena and by deepening their concern about the ever-present "How do we know . . . ? Why do we believe . . . ?" questions.

If one accepts this objective, it is unwise to force a completely rigorous formulation on the students from the very start. Just as it is wise to introduce concepts in kinematics and dynamics by starting with fairly primitive, intuitive ideas, and then to refine and redefine them as one penetrates more deeply into the conceptual structure, so it seems wise to follow a similar pattern in forming the picture of simple resistive circuits. Preliminary concepts of "circuit," "current," "conductor," "nonconductor," "resistance" can be formed initially, in qualitative form, through observations of phenomena in simple resistive circuits. The ideas can then be redefined, quantified, and joined with the concept of potential difference in a more rigorous discussion, which becomes far more intelligible to beginning students than a fully rigorous discussion asserted *ab initio*.

7.2 WHICH SHOULD COME FIRST, STATIC OR CURRENT ELECTRICITY?

There is debate among teachers about this question. Some prefer to introduce static electricity first (as do the majority of textbooks); others prefer to introduce current electricity first and go on to static electricity as a special case afterwards. In this book, static electricity has been discussed first because that is the more common approach. It has always seemed to me, however, that the teacher's own preference is the legitimate determining factor. One can develop the relevant concepts equally soundly and logically either way. What is frequently glossed over without adequate care and attention, however, is the connection between the two sets of phenomena, that is, how do we know that the voltaic battery or the electromagnetic generator maintains continuous transport within conductors of the same "electrical charge" manifest in static frictional phenomena? (This question is discussed in more detail in the following section.)

The fact that the concepts of electricity developed out of electrostatic phenomena historically need not be the pedagogical determinant unless the teacher prefers to follow the historical sequence. Selected portions of the

historical sequence [as treated, for example, by Roller and Roller (1957) in the Harvard Case Histories] can be instructive and illuminating, but more detailed study of the actual historical ups and downs and controversies [as analyzed, for example by Heilbron (1979)] would be hopelessly confusing at an introductory level and would not enhance learning and understanding in most students (however much such study might expand the teacher's own insights.)

Teachers interested in seeing sound developments that start with current rather than static electricity would do well to examine the treatment given by Rogers (1960) and the excellent elementary school unit "Batteries and Bulbs" in Elementary Science Study (ESS) (1968ff). A qualitative treatment for college level students (owing much to ESS) is to be found in Arons (1977) and in McDermott et al. (1996). A continuation that effectively quantifies the qualitative "Batteries and Bulbs" activities is given by Evans (1978).

7.3 HOW DO WE KNOW THAT CURRENT ELECTRICITY IS "CHARGE IN MOTION"?

Many textbooks, after forming the concept of "charge" and examining electrostatic phenomena in the context of frictional electricity, make a discontinuous jump to current electricity by simply asserting that electric circuits containing batteries involve "charge in motion." To most students, however, it is far from obvious, at this early stage, that batteries and household outlets on the one hand, and frictional phenomena on the other, have any connection with each other. They are dumbfounded if the "How do we know?" question is raised explicitly. The visible effects are radically different in the different situations, and the students have not had the benefit of the long series of subtle investigations carried out by Galvani, Volta, and their contemporaries, investigations in which the invention of the voltaic pile came very gradually and out of initially direct connection with electrostatic effects.

Furthermore, the students have, since childhood, heard the words "charge" and "electricity" applied indiscriminately to batteries, household outlets, and electrostatic effects. It has never occurred to them to raise the question as to how one knows that there is any connection among these seemingly disparate systems; the application of the same name from very early on has suppressed the questions that should have been raised. Whether one starts with static or with current electricity in a given course, it is important to stop when the transition is made from one subject to the other and examine the evidence that establishes the commonality asserted in the names.

That this is not a trivial intellectual matter, and that there must have been serious questions about it in the scientific community as late as the 1830s is testified to by the attention given it by Faraday himself. In the *Electrical Researches* [Faraday (1965a)], one finds a paper, dated 1833 and titled "Identity of Electricities Derived from Different Sources," in which Faraday describes the

experiments summarized in Table 7.3.1. The “different electricities” referred to are voltaic, common (frictional), magneto (from electromagnetic induction), thermo (from a thermocouple), and animal electricity from the torpedo and the gymnotus (electric ray and electric eel, respectively). He shows that each of these different sources produces an array of identical effects: Physiological effect (shock), magnetic deflection (of a compass needle as in the Oersted experiment), magnetization of a needle, spark, heating effect (in a conducting wire), chemical action (electrolysis), attraction and repulsion, and discharge by hot air.

Table 7.3.1
 Faraday’s “Table of Experimental Results Showing the Identity of Electricities Derived from Different Sources.” The ×’s denote positive results he obtained himself; the +’s denote positive results filled in somewhat later by other investigators.

	Physiological Effects	Magnetic Deflection	Magnets Made	Spark	Heating Power	Chemical Action	Attraction & Repulsion	Discharge by Hot Air
1. Voltaic Electricity	×	×	×	×	×	×	×	×
2. Common Electricity	×	×	×	×	×	×	×	×
3. Magneto Electricity	×	×	×	×	×	×	×	
4. Thermo Electricity	×	×	+	+	+	+		
5. Animal Electricity	×	×	×	+	+	×		

In the paper accompanying the table, Faraday describes exactly how each of the observations was made. Some of these required the full exertion of his legendary experimental skill. Students are in good company if they initially fail to perceive the identity of the various effects and if they are unable to cite evidence for the interconnections.

“The general conclusion which must, I think, be drawn from this collection of facts,” writes Faraday, “is that electricity, whatever may be its source, is identical in its nature. The phenomena in the five kinds or species quoted differ, not in their character but only in degree; and in that respect vary in proportion to the variable circumstances of quantity and intensity which can be made to change in almost any one of the kinds of electricity, as much as it does between one kind and another.”

Although one certainly need not try to duplicate all of the episodes in Faraday’s table, it is highly desirable to lead students to think about the phenomena involved and to perform a few of the experiments (or at least see them convincingly demonstrated). The most important links to establish at an early stage are probably those among electrostatics, voltaic batteries, and

the household outlet; the connection to other sources will subsequently be plausible without belaboring the issue. Showing Faraday's table and telling the story that goes with it are very helpful in this connection.

Assuming that electrostatic effects have been explored, that the concepts of "charge," "conductor," "nonconductor," "leakage of charge," "electrostatic induction," and "polarization" have been developed, and that the electroscope is now a familiar qualitative device, one can appeal to the following set of experiments with the apparatus sketched in Fig. 7.3.1.

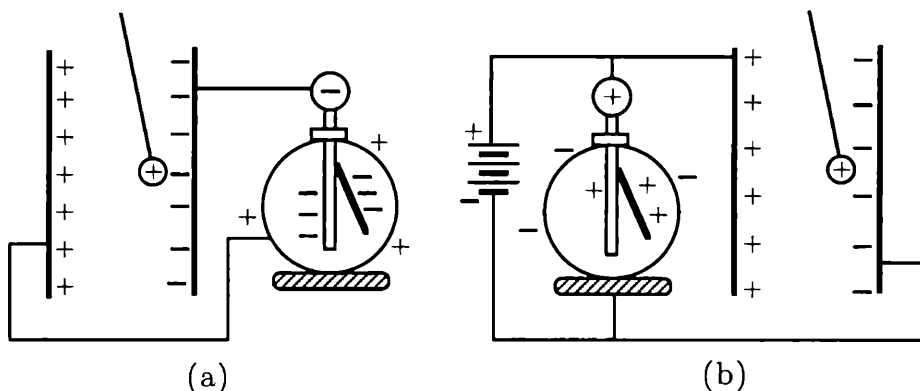


Figure 7.3.1 (a) Capacitor and electroscope charged by induction through contact with charged rod or by connection to Wimshurst machine. Suspended light ball with conducting coating swings back and forth, transporting charge between the plates. System runs down if charge is not resupplied to the plates. (b) Capacitor and electroscope connected to voltaic battery or high-voltage supply running from ordinary electric outlet or electromagnetic generator. Electroscope reveals presence of charge. System behaves exactly as in (a) except that oscillation does not run down.

The capacitor plates can be charged in four ways: (1) By touching one plate with a plastic or glass rod that has been rubbed with fur or silk and touching or grounding the other plate (i.e., charging it by induction); (2) by connecting the plates to the two terminals of a Wimshurst machine or other electrostatic generator; (3) by connecting the plates to the terminals of a voltaic battery of sufficiently high voltage (three or four B batteries of 90 V each usually suffice); (4) by connecting the plates to the terminals of a high-voltage source (rectifier) running off the 120 V line or directly to an electromagnetic generator.

An electroscope is connected across the plates as shown, and deflection of the leaf monitors the charge on the plates. A light ball with a conducting coating is suspended as a pendulum between the plates. When the plates are charged, the ball oscillates back and forth, transporting charge between them. (Students should have seen a demonstration of the oscillating ball in the

electrostatic case and should have been led to sketch and analyze the charge transport prior to the investigation now being undertaken.)

It can now be pointed out that we have constructed a little “motor” (in principle, one could draw mechanical work from the displacement of the ball) in which charge is transported between the plates. Starting with charging mode (1), we note that, as the ball keeps swinging, the deflection of the electroscope decreases: The plates are being discharged; and the swinging “runs down.” We can produce the same final state by connecting a wire (or wet string) between the plates, the discharge taking place so rapidly that we cannot discern a finite time interval, as we can with the swinging ball.

With charging mode (2) we can produce the same swinging of the ball, but the swinging persists, and the electroscope shows deflection, as long as we keep turning the Wimshurst machine. Discharge takes place just as in mode (1) when we cease turning the machine. In other words, the system will maintain continuous transport of charge providing we keep on supplying electrostatic charge to the plates by the frictional process.

With charging modes (3) and (4) the electroscope shows deflection and the ball keeps swinging indefinitely. Since the battery and the rectifier produce the same visible effects (deflection of the electroscope and swinging of the ball) as do the charged rods and the Wimshurst machine, we infer commonality of the underlying electrical effects. (The commonality is, of course, further substantiated by Faraday’s other tests, such as spark, shock, chemical, and magnetic effects.) If a wire (high resistance, of course) or a wet string is used to connect the plates, the ball keeps swinging, and the electroscope continues to show deflection. Since, under these circumstances, transport of charge must be taking place continuously through the wire regardless of whether or not the ball happens to be present, this experiment implies that both of these sources have the capacity of continuously supplying charge to the plates, just as charge is continuously supplied through the frictional process as we keep turning the Wimshurst machine.

The continuous nature of the effects originating in the battery (and also in electromagnetic sources) is further corroborated by the continuous chemical effect observed in electrolysis, by the continuous evolution of heat in wires (the Joule effect), and by the continuous deflection of the compass needle in the Oersted experiment. It is also significant that all of these effects cease immediately when the circuit is interrupted, testifying to the dynamic (as opposed to static) nature of the phenomena.

It is the combination of all of the evidence adduced above that justifies our talking about electrical charge (and its transport) in all of the seemingly unrelated circumstances being considered. Given this common basis and the evidence for continuous transport under circumstances in which a “closed loop” is maintained, we are justified in introducing terms such as “circuit” and “electric current.”

Most textbooks introduce the various technical terms so casually and

abruptly that it seems worth pointing out how Faraday himself felt about the evolving terminology. “Whether there are two fluids or one,” he writes, “or any fluid of electricity, or such a thing as may rightly be called a current, I do not know; still there are well-established electric conditions and effects which the words ‘static,’ or ‘dynamic,’ and ‘current’ are generally employed to express; and with this reservation they express them as well as any other.”

It is worth perceiving and respecting the students’ “reservations” about these matters and giving them an insight into the observations that force us to the concept of continuous transport of electrical charge. Understanding of physics resides in understanding the connection between the phenomena and the concepts as well as in the solving of subsequent problems.

7.4 BATTERIES AND BULBS (I): FORMATION OF BASIC CIRCUIT CONCEPTS

To illustrate the importance of hands-on experience in connection with the formation of the abstract concepts and models we are now discussing, I cite experience with college-age students in connection with electricity and simple circuits. The subjects in this case were preservice (undergraduate) elementary teachers, undergraduate nonscience majors in a general education physics course, and in-service elementary teachers in a summer institute. Somewhere in their school experience or in other circumstances, all had heard about “electric circuits,” most had seen diagrams of electrical configurations in books and on chalk boards, all had been exposed to verbal assertions of facts and concepts of current electricity, even though none had ever taken a physics course.

When these students were given a dry cell, a length of wire, and a flashlight bulb and were asked to get the bulb to light, most started either by (1) holding one end of the wire to one terminal of the cell and holding the bottom of the bulb to the other end of the wire, or by (2) connecting the wire across the terminals (i.e., shorting the cell) and holding the bulb to one terminal. They showed no sense of the functional two-endedness of either the cell or the bulb. Few noticed that the wire became hot when connected across the terminals of the cell, and those who did notice inferred nothing from the observation. It took 20 to 30 minutes for some member of the group to discover, by trial and error, a configuration that lighted the bulb. Then, of course, the message was passed around. Seven-year-old children, incidentally, when given the same task go through exactly the same sequence at very much the same pace.

Absent the synthesis of actual experience into the concept of “electric circuit,” the adults, despite the words they knew, the diagrams or pictures they might have seen, the assertions and descriptions they had read or heard, showed no more understanding of the ideas involved than the seven-year-old approaching the phenomena *de novo*. Purely verbal, passive inculcation had left essentially no trace of knowledge or understanding. In the groups referred

to above, the only individuals who got the bulb lighted quickly were the few who happened to have had previous hands-on experience with electric circuits; one, for example, had been an electronics technician in the service.

Pursuing the sequence thus initiated is one excellent way of leading students to build up the abstract model and concepts associated with simple circuits. They should be led to sketch all the configurations they try out: The ones that do not light the bulb as well as the ones that do. They should be led to separate the two classes and to describe in words what the successful arrangements have in common and how they differ from the other group. Once the intrinsic two-endedness of each object and the necessity of forming the continuous loop have been explicitly recognized, one has arrived at the essential operational description, and the technical term "circuit" can be introduced (note adherence to the precept of "idea first and name afterwards").

Interposing a variety of different objects made of different materials in the loop (coins, pencils, keys, glass, plastic, rubber bands, string, paper, etc.) leads to the classification of materials that allow the bulb to light or prevent it from lighting even when the necessary loop configuration is provided. The operational definition has been formed, and the technical terms "conductor" and "nonconductor" or "insulator" are appropriately introduced. Air can now be identified as a nonconductor.

The situation, now becoming familiar under repeated exploration, strongly suggests the presence of a dynamic rather than a static effect: All effects (heating of the wire, lighting of the bulb) cease when the loop is interrupted or is not completed with a sequence of conducting objects. The two-endedness of each object and the necessity of a complete loop argue for some continuous, dynamic process, as also does the continuous "running down" of the battery. The fact that the bulb lights with equal brightness regardless of where it is placed in the loop (near either terminal, or in between with use of two wires) as well as the fact that the wire, when connected across the battery terminals, is uniformly heated over its entire length, point to continuity and uniformity of the dynamic effect all the way around the system. The uniformity of heating of the wire is supported by the uniformity of heating of stove and toaster elements in the appliances at home (most students turn out to be aware of this uniformity if specifically asked about it but do not think of appealing to the evidence themselves). It must be emphasized explicitly that this uniformity means that whatever effect is taking place invisibly within the conducting loop is not stronger near one terminal than near the other, that is, that no directional effect is detectable.

Evidence for a continuous, dynamic effect, equally strong all the way around the loop, motivates the forming of a mental model for the invisible phenomenon: A continuous flow of some sort around the system. (It must be emphasized that the evidence noted and adduced does not "prove" such a flow "exists" (note Faraday's comment quoted earlier) nor does it tell us what is flowing. What we are doing is forming a plausible picture based on

the qualitative observations, and this picture must be subjected to test and verification through subsequent application in other circumstances). To the continuous flow we now (provisionally) visualize, we give the name “electric current” or simply “current.”

If the concepts of “electrical charge” and “energy” are not yet available from prior study it is, of course, difficult to discriminate what is and what is not being “used up” in an electric circuit. One has no choice but to resort to making some appropriate assertions, perhaps to be substantiated later in the course. If the concepts are available, one can appeal to the prior experience: The law of conservation of charge and the fact that there is no evidence of charge build-up at any points in the circuit support a model in which charge is transported within the system without being “lost” or “emitted” or “transformed.” The emergence of light and heat, however, coupled with “depletion” and chemical change in the battery, or with the supply of mechanical work to a generator, suggest the transformation of energy in association with the continuous transport of (conserved) charge.

One can now exploit the concepts of “circuit,” “conductor,” “nonconductor,” and “current,” by exploring the constitution of the bulb itself as well as that of lamp sockets and switches. **Using the lighting or not lighting of the bulb as an indicator students can be led to explore the construction of each device by identifying the role of conducting and insulating elements.**

Most novices have little or no insight into the structure of the bulb itself. In such instances, it is helpful to break the glass envelope and have them examine the structure carefully. Initially, few students are explicitly aware of the two-endedness of the configuration. They do not distinguish between the metallic tip at the bottom of the bulb and the metallic screw base, nor do they recognize the presence of insulating material and the point and function of the latter. All of these aspects should be traced out by using the lighting or not lighting of a test bulb in order to establish what is connected electrically to what (and what is *not* so connected.) Finally the bulb with the broken envelope can be connected to the battery. The quick flash with which the filament burns out causes astonishment and forms, in itself, an instructive episode.

LED?

Similar investigations need to be carried out with a switch (even if it is only a simple knife switch) and with a socket. To novices such as those who have never yet formed the necessary concepts and do not light the bulb in the first sequence, these investigations and questions are far from trivial. Such students must be led to describe the point and purpose of the structures in their own words (after having identified the locations of the conducting and nonconducting parts) and they must be led to address the questions of what would, or would not, happen if the objects were made entirely of conducting material or entirely of nonconducting material.

Although the latter questions sound absurdly simple, the fact is that they are not. Most students find the entire sequence surprising and illuminating.

It is trivial only to those who have had the prior experience. The latter, of course, should not be subjected to all this detail, but the most common mistake in instruction in the elements of electricity is to assume that these elements are common knowledge and to lose the large number of students for whom this is not the case. A large fraction of students in introductory engineering-physics courses (as many as 30% or more in some instances) are unfamiliar with these basic ideas, and the fraction rises to over 90% in courses for non-science majors. Experience from second grade level on with “Batteries and Bulbs” indicates that the necessary understanding can be readily developed in elementary school through the concrete experience that has been outlined. The understanding is not developed, however, because only an insignificantly small number of elementary teachers have the necessary understanding themselves. Not until we equip our elementary teachers with the security that goes with such understanding will we remove the need for such basic concept building at high school and college levels.

A next logical step in concept development and in refining the model can be the exploration of factors influencing the brightness of the bulb, that is, the intensity of the effect. Connecting a second bulb end-to-end with the first or inserting increasing length of nichrome wire in the circuit both lead to decrease in brightness under the influence of a given battery. This suggests that the “intensity” or “strength” of the electric current might be decreased by inserting more material end-to-end (in “series”) in the loop. Different objects have different degrees of effect. This suggests that different materials and amounts of material are, in some sense, “obstacles” or offer different degrees of “obstruction,” to the flow. From this observation, we form the concept of “electrical resistance.”

Note that the concept emerges naturally and plausibly from the operational sequence described; it is *invented* to fit the observations and the model being induced. In many presentations, the concept is introduced through the assertion that current is decreased “because of the resistance of the material introduced” as though the concept of “resistance” were a necessary a priori and as though the name explained an actual causal mechanism. The student is thus given the impression that “resistance” was something the scientist knew about ahead of time through some process of prior revelation and fails to see it as a concept invented to fit observations one can easily make with everyday objects. I call this mode of presentation “backwards science,” and I urge that it impedes learning and understanding and fails to give the student a clear idea of where scientific concepts come from.

In connection with the concept of resistance (or its inverse, conductance), it is interesting and informative to note how Cavendish compared the conductivities of different metals. Charging a frictional generator with a fixed number of turns and using wires of the same diameter but of different metals, he took the shock of discharging the generator through the wire and through his own body. By comparing the lengths of wire that gave the same intensity of shock,

he arrived at an essentially correct qualitative ordering of the conductivities. [It is instructive to note that Cavendish performed this investigation some 30 or so years before the publication of Ohm's law. Neither current nor potential difference were quantified, but the concepts were being separated nevertheless. And even Ohm, in his original paper, still measures resistance in terms of length of wire (see Section 7.6).] Students can readily see the connection between this story and the invention of the resistance concept in the "Batteries and Bulbs" sequence.

7.5 BATTERIES AND BULBS (II): PHENOMENOLOGY OF SIMPLE CIRCUITS

Research on student learning and concept formation [e.g., Arons (1982); McDermott and Shaffer (1992)] shows that very few students, even among those in engineering-physics courses, develop sound understanding of what happens in simple d.c. circuits through the conventional text presentations based on formulation and application of Ohm's law and Kirchhoff's laws. They may be able to solve conventional end-of-chapter problems, but many are unable to predict qualitatively what will happen to the current at various points in the circuit, or to the potential difference between two specified points, when some change is imposed such as adding or removing a resistive element or shorting two points. This failure indicates serious deficiency in conceptual understanding, but students can be helped to rectify this deficiency if they are given the chance to practice the qualitative reasoning.

Simple configurations of batteries and bulbs lend themselves very nicely to the generation of such practice, and Section 7.13 gives some examples of effective homework and test questions. Many variations on the examples are possible. In connection with questions such as those illustrated in Section 7.13, it is highly desirable for students to perform the actual experiments, and they should be strongly encouraged to do so as part of their homework. The necessary equipment (flashlight bulbs and batteries, sockets, and hookup wire) is readily available in hardware stores and can be readily acquired by the students themselves. In classes where I have had the necessary support, I have maintained a set of kits of such apparatus and have lent the kits to students to take home for one or two nights of homework.

It is desirable to use only one type of bulb in the initial experiments so as to have identical resistive elements in the simple configurations being explored. Different types of bulbs vary greatly in their intrinsic resistance, and, if used in the same circuit, add undesirable complexity to the observed phenomena in the initial stages. (Fortunately, the fact that the identical bulbs have somewhat different resistances when burning at different brightnesses does not significantly affect the qualitative observations and interpretations.)

In connection with such homework, it is desirable to urge students to work

in pairs if they can arrange to do so. The opportunity to talk, argue, and explain in the course of observations and experiments contributes greatly to the learning that takes place. An important caution in connection with such homework: Open-ended assignments, in which the students are given the kits and told to “perform some investigations and experiments,” prove to be virtually useless. The students “mess around” with the equipment, try various arrangements, and so forth, but very few perform any genuine observations or experiments. They fail to notice systematic changes; they do not impose systematic alterations on a configuration and predict or interpret the resulting effects; they fail to invent interesting and fruitful configurations of their own—at least in the initial stages. The majority of students have to be *guided* into a series of investigations by being supplied with some initial suggestions and leading questions (not cookbook instructions that destroy all the inquiry). Many then proceed to develop fruitful sequences of their own, but some never achieve this level of inquiry. [One example of the kind of guidance and questioning that many students need is to be found in Chapter 9 of *The Various Language*, Arons (1977)].

There are several very basic aspects of simple resistive circuits to which students should be explicitly exposed in the initial stages of such investigations in order to counter various naive ideas and preconceptions:

- (a) It is quite plausible to virtually all students that the total effective resistance of a circuit increases as materials (or bulbs) are inserted end-to-end (or in series). It is far from plausible, however, to many students that the effective resistance of a combination of resistive elements decreases as more elements are added in parallel. One is “adding more resistance,” and the very robustly held naive notion is that the “resistance must increase.” Having studied Ohm’s law and developed the formulas for series and parallel combinations does not remove this preconception except in a few of the quicker students. It is helpful to lead students into performing batteries-and-bulbs experiments that concretely show the lowering of effective resistance in parallel combinations: A combination of several bulbs in parallel runs a battery down more quickly than does one bulb alone; starting with two bulbs in series, the brightness of the one bulb is increased as additional bulbs are added in parallel with the other bulb; if a short thread of steel wool is inserted in series with a bulb and additional bulbs are then added in parallel with the first one, the steel wool burns out at some point in the sequence; and so on. (The last experiment is one the students usually enjoy, and it also, of course, leads into a discussion of the point and purpose of fuses and circuit breakers.)
- (b) When two bulbs are connected in series and a wire is then connected across one of the bulbs (short circuit), many students (even among those in an engineering-physics course) are astonished by the fact that the shorted bulb goes out and the other burns more brightly. They have

all heard the term “short circuit” but very few have any operational or conceptual awareness of its meaning, nor do they visualize the accompanying effects. All they know is that a short circuit is something “bad.” They should be led into forming the concept through “idea first and name afterwards.”

- (c) Readers will have noticed that the initial experiments being suggested have, at least by implication, been confined to the use of a single battery and to the exploration of the effect of distribution of resistive elements on the current in a circuit, that is, to only the first steps in the elaboration of the current model. A next step, of course, involves the perception that still another factor is involved in determining current anywhere in a circuit, namely a property of the battery itself. A qualitative concept of “strength” of a battery can be introduced through experiments with combinations of batteries in series (both “adding up” and “subtracting” their effects). This is, of course, a stepping stone to the formation of the concept of “potential difference,” but it is not desirable to plunge into the latter concept precipitously. The first step is to show that current is determined by an internal property of the battery as well as by the distribution of resistance in the external circuit. Formation of the concept of “potential difference” is a subtle enterprise in its own right.

7.6 THE HISTORICAL DEVELOPMENT OF OHM’S LAW

Although I do not advocate the historical sequence above other approaches, I find it illuminating to be aware of how Ohm’s law developed and evolved because the steps of evolution reveal how the various subtle insights were originally achieved. Such awareness helps one appreciate difficulties experienced by the students when modern shortcuts are imposed.

Ohm’s work, published in 1826 [see translated excerpt in Magie (1935)], did not lead directly to the relation we now call Ohm’s law. His investigation was, in a sense, a quantitative version of Cavendish’s qualitative comparison of the shock sensed by one’s body in taking the discharge from an electrostatic generator through different lengths of metallic wire (see Section 7.4.)

Ohm initially attempted to make precise measurements of current supplied, by voltaic batteries with different numbers of plates, to metallic wires of different length and different materials. He made use of a very carefully designed tangent galvanometer (the device had been invented in 1820 as an immediate consequence of the electromagnetic discoveries of Oersted and Ampere), but he found it impossible to obtain reproducible measurements because of erratic fluctuations in the voltaic piles. (It was a long time before the effects of temperature and pressure, cleanliness of surfaces, uniformity of electrolytes, polarization of the electrodes, etc., were brought under control to the extent

of producing reasonably stable voltaic batteries.) In consequence, at the suggestion of Poggendorf, he turned to the thermoelectric effect, which had been discovered by Seebeck in 1822.

Ohm constructed a thermocouple consisting of a bismuth-copper junction with the ends maintained at different temperatures. The ends of the thermocouple were connected to cups of mercury, and the wire being investigated had one end immersed in each of the mercury cups. The galvanometer had a torsion suspension made of a thin ribbon of gold leaf (which was shown to be highly elastic in its return to zero), and the steel compass needle suspended at the end of the gold ribbon was carefully held at a fixed distance above one of the rigid conductors running from one end of the thermocouple to one of the mercury cups.

Ohm recorded the torque on the tangent galvanometer as a function of the length of wire inserted (the diameter of the wires was fixed) and found the empirical relation

$$\tau = \frac{a}{b+x} \quad (7.6.1)$$

where τ denotes the observed torque, a is a constant for a given temperature difference between the ends of the thermocouple, b is a constant property of the "base" (consisting of the thermocouple and all the connections, including the mercury cups), and x is the length of wire being inserted.

Ohm interpreted a as a measure of the "exciting force" driving the electric current and eventually showed it to be directly proportional to the temperature difference between the two ends of the thermocouple (note that no connection is made, at this point, to the electrostatic concept of "potential difference"). He also, as is to be expected, interpreted the quantity $b+x$ as the resistance of the entire circuit with b as the contribution due to the frame (in units of length of wire). He thus shows resistance of wire of a given material to be a linear function of length if diameter is held constant. He also showed that resistance increases as the temperature of the wire increases.

Ohm went on to show that "an inch of brass wire is equivalent to 20.5 inches of copper wire" and, on using a very long brass wire resulting in a barely detectable current, remarks "we see . . . that the equation fits with experiment very accurately nearly up to the extinction of the force by the resistance of the conductors."

It should be noted, however, that Eq. 7.6.1 is not yet "Ohm's law" as we usually state the latter in our texts. As indicated earlier, it is much more nearly a quantified version of Cavendish's investigation.

The next step toward the modern version of Ohm's law was taken by Joule in 1841 through the experiment shown schematically in Fig. 7.6.1. Joule wound different lengths of wire (the *ratio* of resistances is thus known) on glass tubes, immersed the windings in cylinders of water, and connected them, in series with each other, to a voltaic battery. He showed that the rate of

evolution of heat in each cylinder was in the same ratio as the resistance of the wires, and he observed that the rate of evolution of heat in each cylinder was proportional to the square of the current as measured by the tangent galvanometer, that is, he showed that the rate of evolution of heat in a current-carrying wire is proportional to the product I^2R . It is from these experiments and conclusions that the terms “Joule’s law” and “Joule heat” derive. (It is interesting to note that Joule fully anticipated the relation to current: “I thought the effect produced by the increase of the intensity of the electric current would be as the square of that element . . . arising from increase of the quantity of electricity passed in a given time and also from the increase of the velocity of the same.”)

During the 20 years that followed Ohm’s experiments of 1826–27, his results were repeated and confirmed with increasing precision by a number of investigators, but the connection between the law of conduction and other electrical concepts remained obscure until the essential unity of all the various phenomena was grasped through growing appreciation of the law of conservation of energy.

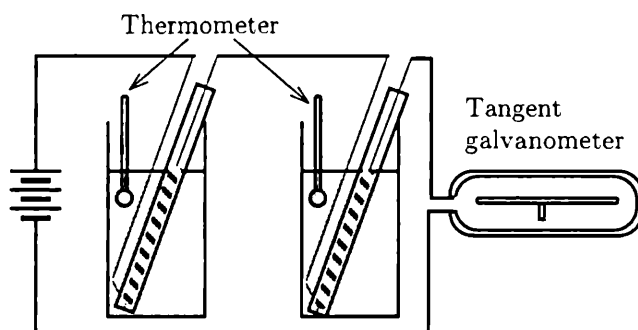


Figure 7.6.1 Schematic diagram of Joule’s apparatus for the determination of the evolution of heat in current-carrying wires. Different lengths of resistive wire of the same diameter are wound on glass tubes and immersed in cylinders of water. The wires are connected in series through a battery and a tangent galvanometer. Thermometers measure the temperature changes of the water, and current is measured by the galvanometer.

Roget as early as 1832 and Faraday in 1840 had both pointed to the chemical changes taking place within voltaic batteries as evidence that something was being “used up” within the system while electrical charge was being displaced. In Roget’s words, “All the powers and sources of motion, with the operation of which we are acquainted, when producing their peculiar effects, are expended in the same proportion as those effects produced; and hence arises the impossibility of obtaining by their agency a perpetual effect or, in other words, a perpetual motion.” And from Faraday came the remark that one never found in nature “a pure creation of force; a production of power

without a corresponding exhaustion of something to supply it.” (It is interesting to note that both of these statements preceded the papers of Mayer and Joule in the 1840s. The idea of conservation of energy was “in the air” as a qualitative “postulate of impotence,” albeit not fully formulated or substantiated quantitatively.)

In 1849, just as the accumulated impact of the conservation of energy concepts promulgated by Mayer, Joule, and Helmholtz was being felt by the scientific community, the German physicist Kirchhoff pointed out that existing knowledge of electricity fell together beautifully and consistently if the quantity a in Ohm's empirical relation (Eq. 7.6.1) were identified with the magnitude of the potential difference imposed between the ends of a conductor of resistance $R = b + x$. This finally put the relation in the form familiar to us, namely:

$$I = \frac{\Delta V}{R} \quad (7.6.2)$$

The identification of a with ΔV accounted for the previously established fact that the difference of electric “tensions” between the terminals of a voltaic battery (as measured by means of an electrometer) was proportional to Ohm's quantity a . Furthermore, since the rate P at which work is being done in the displacement of charge must then be given by

$$P = I\Delta V \quad (7.6.3)$$

it follows, by combination with Eq. 7.6.2, that

$$P = I^2 R \quad (7.6.4)$$

for a resistive element.

The subtle logic of these insights should be carefully noted: There is no a priori reason why all the work done in displacing electrical charge should be converted into thermal internal energy of the system. In fact, there are many instances (e.g., the electric motor) where this is not what happens. In the purely resistive circuit, however, *all* the electrical work supplied is converted into thermal internal energy.

This is what is confirmed by Joule's experiment and is expressed in “Joule's law”; it cannot be “derived” a priori, and this is why Joule's law should be recognized as having an essentially independent status, establishing the fact that electrical work is completely converted into thermal energy in a purely resistive circuit. Most textbook presentations simply combine Eqs. 7.6.2 and 7.6.3 without explicit justification.

Because of the way in which most textbooks introduce resistive circuits and concentrate problems and exercises on manipulation of Ohm's law and its consequences, very few students absorb awareness of the fact that current in a system is *not* always determined by Eq. 7.6.2. It might be determined

independently by the power requirement of Eq. 7.6.3, and, in that case, Eq. 7.6.4 is not applicable. (This conceptual aspect is discussed more fully in Section 7.11, where it is suggested that examining the question of why power transmission at high voltages is significantly more efficient than transmission at lower voltages helps students acquire a better understanding of when these relations are applicable and when they are inapplicable.)

7.7 TEACHING ELECTRICAL RESISTANCE AND OHM'S LAW

The teaching of Ohm's law varies widely among present-day textbooks. Many show at least some care in erecting a motivated and logical conceptual structure intelligible to the level of student addressed, but many assert the concepts of "current" and "resistance" as though they were simple primitives that require no development or explanation. Very few justify the application of the concept of "potential difference,"² usually developed in connection with electrostatic effects, to circuits; they simply assert the applicability as though it were obvious to anyone. (Witness how long it took nineteenth-century physicists to perceive the interconnection.)

Among the more careful modes, it seems to me that no single one emerges at the present time as dominant or pedagogically superior, and teachers must, in any case, accommodate themselves as best they can to the approach in the textbook they have selected. In most instances, the concepts and relations are asserted more or less *a priori* and are eventually justified by the fact that they "work," that is, the results agree with observations. Unfortunately, this *ex post facto* kind of approach is unsatisfactory to many good students. The logic is unclear to them since it is not explicitly outlined and, if given the chance, they much prefer to see how something came to be known than to have it asserted *ex cathedra* and then shown to be valid. In the former circumstances, they have the feeling that, in principle, given time and opportunity, they may have formed the ideas on their own; in the latter, *ex cathedra*, case, they feel they could never have possibly come up with the ideas. And in the latter feeling, they are on sound ground because the ideas were never, in fact, arrived at in that way.

Iona (1979) gives an excellent survey and commentary on the existing situation in the textbooks. It should be read for its own sake and will not be

²It should be noted at this point that the very widely used term "voltage" generates more confusion than enlightenment. The term is used differently by different authors in different contexts, and no clear, systematic operational meaning has crystallized. The result of this variability is to blur, for many students, the distinction between emf and potential difference. The usage also tends to give naive learners the false impression that "voltage" is a property of a circuit at one single location. It is necessary to have the students use the term "potential *difference*" over and over again to fix the insight that *two* points are always involved. It is for this reason also that I systematically use the symbol ΔV rather than the symbol V throughout this discussion.

incorporated here except for the pointing up of a few elements. Iona found that, in the textbooks he examined, the most prevalent approach was to start with the ratio of potential difference to current as the defining expression for resistance, that is

$$R \equiv \frac{\Delta V}{I} \quad (7.7.1)$$

This is a legitimate approach, especially if the students have been helped to acquire a preliminary, qualitative, nonrigorous notion of “resistance” such as that advocated in Sections 7.4 and 7.5. Equation 7.7.1 then becomes a redefinition and quantification of the concept.

When the consequences of the definition in Eq. 7.7.1 are explored for different conducting materials and systems, it is found that, in simple metallic conduction, R is constant in the sense of being independent of ΔV and of I , although it is, in general, a function of the state (temperature and pressure) of the material. Systems for which R is constant, and for which the relation between ΔV and I is therefore linear, are said to obey Ohm’s law. Systems for which the relation is not linear are said to be “non-Ohmic” or not obeying Ohm’s law, but the definition of R , now as a variable property, still holds.

An alternative approach appearing in some texts is to start with Joule’s law and define resistance as the ratio of power supplied to the square of the current. This, of course, involves prior understanding of the energy concepts and is even more abstract than the approach via Ohm’s law. Connecting it to Ohm’s law also involves the a priori assumption that all of the work done by the source is converted into thermal internal energy, and this tends to prevent the student from clearly separating the elements of the logical structure.

One very fundamental and simple implication of Ohm’s law (as it applies to metallic conduction) is rarely made explicit to the students, yet it illustrates what deep inferences can sometimes be drawn if one recognizes what is *not* the case as well as what is. The fact that the straight line extends all the way down to zero and shows no evidence of a measurable voltage threshold for the onset of current implies that the charge carriers, whatever they may be, are free and unbound in the structure of the metal. (Dielectric breakdown in nonconductors, on the other hand, requires the application of very high potential differences before the bound charge carriers are freed and significant conduction begins.) This is a kind of physical insight students should be helped to acquire in addition to practice at application of Ohm’s law itself.

7.8 IS ELECTRIC CURRENT IN METALS A BULK OR SURFACE PHENOMENON?

Very few textbooks pay attention to the question raised in the title to this section, yet it is a question that involves important physics and phenomenology,

and consideration of this question can significantly enhance student understanding. We do know that transport of electrical charge in metallic circuits takes place, under ordinary circumstances, through the body of the metal rather than along the surface, but how do we come to know this? (That this is not a trivial question is indicated by the fact that, at high frequencies, the so-called “skin effect” comes into play, and under these circumstances conduction is confined to the surface layers.)

Most textbooks simply assume that conduction is a body effect and do not even inform the student that an assumption is being made. The more advanced textbooks take current density J to be uniform over the cross section of a conductor and assert the relation

$$\vec{E} = \rho \vec{J} \quad (7.8.1)$$

where \vec{E} denotes the electrical field strength within the body of the conductor and ρ the resistivity of the material. They then go on to show, by invoking Ohm’s law, that the total resistance R will be given by

$$R = \rho \frac{l}{A} \quad (7.8.2)$$

where l and A denote the length and cross-sectional area of the conductor, respectively.

But what justification is there for assuming bulk flow in the first place? Students should be helped to raise this question and to consider how it might be resolved. The strongest primitive testimony to bulk rather than surface conduction resides, of course, in the *empirical* fact that the total resistance R is found to be directly proportional to the length of a wire and inversely proportional to the cross-sectional *area*, that is, to the *square* of the cross-sectional dimension:

$$R \propto \frac{l}{A} \propto \frac{l}{D^2} \quad (7.8.3)$$

The inverse proportionality to square of cross-sectional dimension is what speaks to bulk conduction. Had the resistance been inversely proportional to the *first power* of the diameter of the wire instead of the *square* of the diameter, the indication would have been surface conduction. Here the student must confront thinking about what is *not* as well as what *is* the case, and such thinking is deeply conducive to understanding (and to cognitive development as well). Furthermore, this episode exhibits the power and importance of ratio reasoning in circumstances where numbers are not involved but a significant *conceptual* insight is attained from the *functional* relationship.

7.9 BUILDING THE CURRENT-CIRCUIT MODEL

Our picture of electric current, even in “simple” resistive circuits, is very abstract, subtle, and sophisticated. It is cryptically asserted in textbooks with

very little discussion of observation of effects—the more elementary the text the more cryptic the assertion—and then teachers wonder why students have wildly erroneous ideas concerning what is happening.

An initial, nonrigorous approach, requiring subsequent elaboration and redefinition, is outlined in Sections 7.4 and 7.5. The closing of a still open question is outlined in Section 7.8. There remain additional questions: Is one variety of charge being displaced or both? Is charge supplied only by the source or is the conductor itself a container of displaceable charge? What are the respective roles of potential difference and resistance in establishing currents in various configurations? Can we visualize what is happening in an electric circuit by appealing to analogy with some more intuitively comprehensible mechanical system (e.g., the fluid model)?

Some of these questions are widely ignored or are treated by assertion without addressing the underlying “How do we know . . . ?” questions. Others are carelessly handled. As shown, for example, by Cohen, Eylon, and Ganiel (1983), many students emerge with the conception that potential difference is a consequence of displacement of charge rather than its cause.

Several pilot attempts to evolve more carefully logical and intelligible presentations have been published [e.g., Evans (1978); Steinberg (1983)], but further polishing and work with students are needed. Steinberg, in particular, exploits a batteries-and-bulbs situation combined with inexpensive capacitors of very high capacitance. With these capacitors in the circuits, one can observe transient lighting of the bulbs as the capacitors are charged and discharged. This additional ingredient enormously extends the number of questions that can be asked and answered about location and displacement of charge and greatly enhances the soundness and completeness with which the current model can be constructed. Steinberg outlines a sequence of investigations in which fluid models (including that of a compressible fluid) are tested, found wanting in certain respects, and lead to a final picture of conduction that is fully correct in its own right.

Steinberg’s approach makes it possible to attack a common student misconception that circuit elements affect the situation “downstream” while not affecting anything “upstream.” Furthermore, the steady-state situation in continuous circuits conceals temporal effects that are important in understanding construction of the electric current model, and continuous circuits do not reveal, as clearly, charge displacements associated with various potential differences.

Teachers would do well to study some of these developments with a view to incorporating into their own teaching those aspects they find congenial and helpful to the students.

7.10 CONVENTIONAL CURRENT VERSUS ELECTRON CURRENT

Especially at precollege, but also at college level, many textbooks are asserting that electric current consists of a flow of electrons. Some texts do not even bother making it clear that they are talking only of metallic conduction. The electron concept is asserted without evidence or basis for acceptance, and students are left with no frame of reference in which to set the term: They are left with no sense of how the properties of this entity (the electron) compare with those of any other on the microscopic scale, how we come to know anything about it in the first place, or what role it plays in the structure of matter in general.³ All they have is a name (i.e., jargon), and this undermines their capacity to distinguish between jargon and knowledge.

The excuse frequently given for this approach is that the electron picture is “correct” and that students should be given the correct answers as scientists know them. The result is that students are misled in many ways, and they are also robbed of a valuable opportunity for phenomenological thinking and reasoning that enhances understanding and insight not only into this particular concept but into how science works in general.

When electron current is asserted in the manner described above, most students end up with the completely false notion that all conduction consists of displacement of electrons and hence of negative charge. They end up with this misapprehension even if it has been qualified to apply only to metals. (This is due to the fact that they are not given the opportunity to encounter any cases of conduction other than in metals, and the restriction is lost sight of and forgotten even if it was ever noticed.) In fact, of course, metallic conduction is a fairly special case, accompanied only by the relatively esoteric instances of conduction under photo or thermal emission. In most electrolysis, in ionized gases, in dielectric breakdown, conduction takes place by migration of both positive and negative ions simultaneously. In some cases of electrolysis (e.g., electroplating of copper or silver), conduction takes place by migration of positive ions only. In semiconductors, it is in some circumstances more convenient to visualize the migration of positive “holes” than of electrons. Thus the premature assertion of electron current tends to plant a narrow misconception in student minds rather than liberate them with a “right” answer.

I suggest that it is pedagogically wiser and more fruitful to teach students the positive current convention, and I join Iona (1983) in his forceful argument to this effect. In addition to the fact that only normal metallic conduction and photo and thermal emission are purely electronic, there are a number of other reasons that favor adherence to the positive current convention.

³A few textbooks (*PSSC Physics* for example) try to provide a sound basis for accepting the electron concept before asserting the electron picture of conduction, but these are rare, and unsubstantiated assertion is the norm.

Students, in their initial exposure to electric circuits, tend to be extremely reluctant to accept the fact that it is impossible to determine, through ordinary macroscopic observations (either electrical or electromagnetic), whether it is negative charge, positive charge, or both varieties that are being displaced in metallic conduction. Observed effects can be equally well accounted for and predicted on the basis of any one of the models. It is only the Hall effect [see, for example, Magie (1935); Resnick and Halliday (1985)] that indicates transport of negative charge in ordinary metallic conduction, and it is unable to say anything about the carrier. Just as in electrostatics, students should confront the indeterminacy as to whether positive charge, negative charge, or both are being displaced; they should first see the same indeterminacy in metallic conduction and recognize that the positive current convention is, in this sense, as good as any other. It is only such experience that registers the insight that positive current is a convention and not a fact. This helps the students toward far deeper appreciation of the electron picture when the latter is finally substantiated and makes them far more sensitive to the “How do we know . . . ? Why do we believe . . . ?” questions. They see the course of question asking and development and not just the sterile end result.

The principal reason for retaining the positive current convention is that it underlies: (1) the definitions of electric field strength and potential difference; (2) the treatment of capacitive and inductive circuit elements; (3) all the standard mnemonics of electromagnetism and the Maxwell equations; and (4) the standard notation in diagrams of electronic circuits. This convention has not been, and will not be, changed because of the acceptance of electron current. Students who come to electromagnetism without having had experience of the positive current convention suffer severe and unnecessary confusion without having had any real pedagogical gain.

7.11 NOT EVERY LOAD OBEYS OHM'S LAW

Given the usual presentations of current electricity, very few students, even in engineering-physics courses, emerge with the realization that Ohm's law does not apply in all circumstances in which a source of potential difference drives an electric current. Just telling the students this, of course, leads nowhere. (I am speaking here not of nonlinearity in resistance but of loads that are not purely resistive, i.e., running a motor, charging a battery.) An effective way of leading students to confront this issue is to raise the question of why it is that power transmission is far more efficient at high voltages than at low. This embeds the idea in a rich and memorable context and gives it a connection to engineering and societal questions.

Even though power transmission is alternating current (a.c.) rather than direct current (d.c.), there is no serious error in treating this as a d.c. problem at an elementary level. Suppose we wish to supply a fixed amount of power

P_L at a potential difference ΔV_L to a given load. The resistance of the transmission lines from the distant source is denoted by R_T . It should be specified that the nature of the load is not known and that the voltage level is to be chosen on the basis of investigation of the favorable conditions.

The crux of the matter is that the resistance of the load is irrelevant even if it is known. The current drawn from the supply lines is determined by the fixed power requirement:

$$I = \frac{P_L}{\Delta V_L} \quad (7.11.1)$$

regardless of whether or not the load is resistive. This is what students should be led to see, and it is usually a significant obstacle in the minds of many. What they want to do, by habit, is apply Ohm's law.

Since the current is fixed by the power requirement, the Joule heat loss in the power lines is now fixed:

$$\text{Resistive energy loss} = I^2 R_T \quad (7.11.2)$$

The efficiency η is given by

$$\eta = \frac{P_L}{P_L + I^2 R_T} \quad (7.11.3)$$

Combining Eqs. 7.11.3 and 7.11.1 gives

$$\eta = \frac{1}{1 + \frac{P_L R_T}{\Delta V_L^2}} \quad (7.11.4)$$

Examination of Eq. 7.11.4 shows that the efficiency rises dramatically as the potential difference is increased, and this is the basic reason for construction of very high-voltage transmission lines. Students should be led to explore and interpret this equation both graphically and numerically. In Rogers (1960), one of the few textbooks dealing with this illustration at an elementary level, students are led through a good numerical comparison of two cases without the algebraic analysis.

Whether this problem is examined algebraically or arithmetically, it is important to lead students through the reasoning Socratically, having them fill in and explain the steps and interpret the results. Presenting them with the derived result and calling for some algebraic analysis or numerical substitution does not register an understanding of the crucial aspects of the reasoning.

7.12 FREE ELECTRONS IN METALS: THE TOLMAN-STEWART EXPERIMENT

In many textbook discussions of metallic conduction at secondary school and introductory college levels, the existence of electrons and their role as free entities in metals are simply asserted on the "scientists know that . . ." basis.

Given a large volume of such asserted end results, many students, especially the slower ones who are still developing abstract reasoning capability, are unable to discriminate what, of “knowledge” they “possess,” is based on evidence and understanding and what consists of memorized, unsupported assertions. This is not a healthy intellectual condition, and, for those students who will not become scientists and to whom such study is a matter of general education, this condition is destructive of any understanding of the nature, power, and limitations of science. With such background, they do not become citizens who comprehend the nature and role of science in intellectual history and in our society. I have long voiced my opposition to such presentations of physics, and I continue to do so.

If one is to discuss the role of electrons in metallic conduction (and this can perfectly well be done, if desired, while still teaching the positive current convention), this can and should be done by first laying some basis for understanding why we believe in electrons, when and how they are detected, what their properties are compared to those of other entities on macro- and microscopic scales. Some textbooks that deal with electrons do lay such a basis in a plausible and reasonable, even if not fully rigorous, way, and I consider this fair game. It is the bald, unsubstantiated assertion that is destructive.

Given some plausible basis for understanding the evidence underlying acceptance of the electron concept, the next “How do we know . . . ?” question has to do with the justification for believing electrons to be the free carriers of charge in metals. Freedom of the charge carriers is inferred from the observed fact of zero threshold potential difference for the initiation of current in the metallic conductor, that is, from the fact that the straight line in Ohm’s law extends all the way to the origin (see Section 7.7). As often happens, the historical sequence is very illuminating in this respect, and, at the same time, allows the kind of spiralling back in later, richer context that has been advocated throughout this book.

After Thomson’s initial (1897) work on the corpuscular nature of the cathode beam (see Sections 10.4 and 10.5), he proceeded to show that particles with the same charge-to-mass ratio appeared outside the metal in both photoelectric and thermal emission. [The cathode beams he studied (in Crookes’s tubes with considerably improved vacuum) were produced, we now understand, through bombardment of the cathode by positive ions formed in the residual gas still present.] These observations led naturally to the widely held supposition that electrons were present as such within metals and were probably the free carriers of charge. Supposition, however reasonable, is nevertheless not evidence. There remained the problem of showing that unbound electrons were indeed present, as such, within metals. This problem was attacked in a very direct and simple way by Tolman and Stewart (1916), (1917).

The procedure adopted by Tolman and Stewart was that of imparting a large acceleration to a metal sample and observing the effect of any potential difference set up between the ends by virtue of the presence within the metal

of free charge carriers possessing inertial mass. Given the presence of such carriers, the effect would be analogous to that in a closed, accelerated tube of fluid. In the fluid, a pressure gradient would be set up, the hydrostatic pressure becoming higher at one end than at the other, until all the fluid within the tube had the same acceleration. In the metal, the charge carriers would shift within the lattice, becoming more concentrated at one end than at the other until the electrical field thus set up within the body of the metal would be sufficient to accelerate all the free charge carriers along with the accelerated lattice.

In the actual experiment, Tolman and Stewart used a rapidly rotating spool (diameter about 24 cm) carrying about 450 m of wire, connected to a very sensitive ballistic galvanometer. The rotation of the spool was stopped abruptly, and deflection of the galvanometer revealed the passage of a pulse of current. From this measurement, it was possible to infer the charge-to-mass ratio of the free carriers within the metal, and, within fairly large experimental uncertainty, this ratio was equal to that observed by Thomson and others for electrons outside the metal. At an introductory level, however, complete analysis of this measurement, including interpretation of the behavior of the ballistic galvanometer, would tend to cloud the main issue and is unnecessary. (For more advanced students, in engineering-physics courses for example, the full analysis can be very instructive because of the richness of physical ideas brought together in a significant context—even though the era of the ballistic galvanometer is long past and is remembered only by a vanishing generation.)

Assuming that one can measure, directly, the imparted acceleration and the potential difference set up between the ends of the accelerated conductor, the analysis becomes exceedingly simple, as illustrated in Fig. 7.12.1:

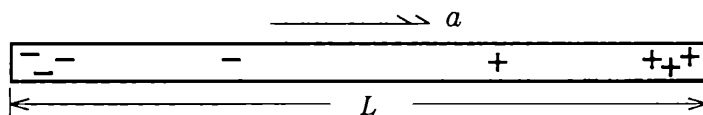


Figure 7.12.1 Schematic diagram of Tolman-Stewart experiment: Metallic conductor of length L is accelerated to the right at known acceleration a . If free charge carriers possessing inertial mass m and charge e are present within the body of the conductor, they will tend to shift, becoming more concentrated at the left than at the right end of the conductor. The result will be a potential difference ΔV between the ends and a field strength $E = \Delta V/L$ within the body of the conductor.

Applying Newton's second law to the individual charge carriers, we have

$$Ee = ma \quad (7.12.1)$$

and

$$\frac{e}{m} = \frac{aL}{\Delta V} \quad (7.12.2)$$

Thus the unknown charge-to-mass ratio is expressed in terms of measurable quantities. In the actual experiment, a potential difference does indeed develop between the ends of the conductor, and the calculated charge-to-mass ratio agrees with that obtained for electrons ejected from the metal. Tolman and Stewart first performed the experiment with copper wire and subsequently showed that the same results were obtained with silver and aluminum. The simplified analysis outlined above is highly instructive for all students because of the opportunity it affords for spiralling back to the use of earlier concepts in setting up the present problem. Furthermore, the context is not that of an artificial end-of-chapter exercise but a real experiment with a significant interpretation.

My colleague (the late) Philip Peters informed me that he tried this out as a problem for *graduate* students in an advanced electricity and magnetism course and found that the students had great difficulty visualizing the situation and dealing with the physics. Such response from the students stems, of course, not from the intrinsic difficulty of the problem, but from the fact that they never had a chance to think about such situations—situations that they should have had opportunity to consider and visualize in undergraduate work.

Students should be led to see the connection between this electrical problem and such mechanical analogs as the pendulum suspended in the accelerating car and the ball on the cart being accelerated by the rear wall of the accelerating cart. A more closely analogous mechanical situation is that of fluid in an accelerated tube—the acceleration being either linear or centripetal.

In the case of an essentially incompressible fluid of unknown specific volume v in an accelerated tube of length L , the analogous result is

$$v = \frac{aL}{\Delta p} \quad (7.12.3)$$

where Δp is the measured pressure difference between the ends of the tube and a is the acceleration imparted.

The derivation of this result is an excellent exercise for honors students. Not only does it point up the deep similarity between apparently unrelated physical situations, but it also sets up the basis for understanding, in the light of the pressure gradient developed, how a cork with lower density than the surrounding fluid comes to be displaced in the same direction as the acceleration rather than in the opposite direction.

A more sophisticated, but still analogous, problem is that of a tube with closed ends containing an ideal gas at initially uniform pressure p_0 . When the tube is accelerated, a pressure gradient develops as in the isothermal gas in a gravitational field. One must integrate the differential equation for the pressure, obtaining an exponential as in the case of the isothermal atmosphere, and one must formulate and apply the condition for conservation of mass in the tube. For this situation one could, in principle, find the molecular mass μ

of the gas from the measurable quantities since one obtains:

$$\mu = \frac{RT}{aL} \frac{\Delta p}{p_0} \quad (7.12.4)$$

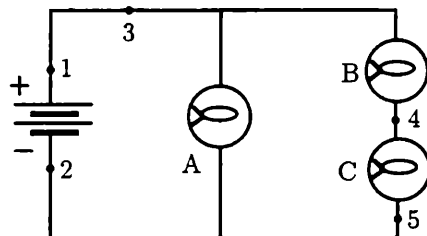
where Δp denotes the pressure difference between the ends of the tube. Although this is certainly not a useful way of measuring molecular mass, able students find the exercise in mathematical physics very instructive.

7.13 SAMPLE HOMEWORK AND TEST QUESTIONS

NOTE: Following are several sample problems on qualitative, phenomenological aspects of simple d.c. circuits—aspects seriously neglected in most textbooks. Such problems are useful as homework or as open-book test questions. These are simply prototypes; it is obvious that many variations are possible (and desirable.) Exposure to *several* such questions is necessary before the majority of students become successful in the reasoning.

1 In the circuit shown in the following figure, the battery maintains a constant potential difference between its terminals at points 1 and 2 (i.e., the internal resistance of the battery is to be considered negligible). Three identical flashlight bulbs A, B, and C are screwed into their sockets and are lighted when the circuit is closed.

After each of the changes suggested in the following questions, the system is returned to the initial condition shown in the figure before the next change is made. (The question “What happens to . . . ?” refers to whether the quantity in question increases, decreases, or remains unchanged. Indicate your reasoning briefly in answering each question.)

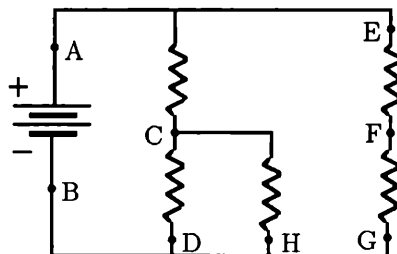


- How do the brightnesses of bulbs A, B, and C compare with each other in the initial condition?
- What happens to the brightness of *each* of the three bulbs when bulb A is unscrewed and removed from its socket? What simultaneously happens to the current at points 3, 4, and 5?
- What happens to the brightness of *each* of the three bulbs when bulb C is unscrewed and removed? What simultaneously happens to the current at points 3, 4, and 5?
- What happens to the brightness of *each* of the three bulbs if a wire is connected from the battery terminal at point 1 to point 4? What simultaneously happens to the current at point 3? What simultaneously happens to the potential difference across bulb B? What simultaneously happens to the potential difference across bulb C? What happens simultaneously to the potential difference between points 1 and 5?

- (e) What happens to the brightness of *each* bulb and to the current at point 2 if a wire is connected from the battery terminal at point 2 to the socket terminal at point 5?
- (f) What happens to the brightness of *each* bulb if a fourth bulb (D) is connected in parallel with bulb B alone, i.e., not in parallel with both B and C? What happens simultaneously to the current at point 3? What happens simultaneously to the potential difference between points 3 and 4? To the potential difference between points 4 and 2?

NOTE: The following question is a version of the preceding type in a multiple choice format. Since it allows rapid correcting, it is useful as a test item after questions of the preceding form have been used in sufficient number. Without the requirement of verbal explanation, however, the right answer is still not necessarily indicative of genuine understanding.

2 Consider the resistive circuit shown in the adjacent figure. The battery has negligible internal resistance, and the resistors are all identical. Circle the correct choice of word or words in the statements following each suggested change in the circuit.



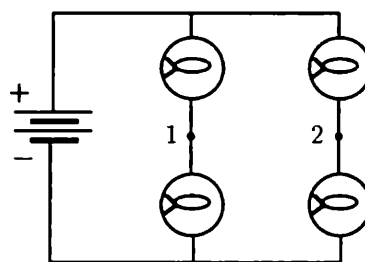
- (a) Another identical resistor is connected between points E and G.
- (1) The current at point A (increases) (decreases) (remains unchanged).
 - (2) The current at point F (increases) (decreases) (remains unchanged).
 - (3) The potential difference between points E and F (increases) (decreases) (remains unchanged).
- (b) Return to the initial conditions shown in the figure. An identical resistor is connected between points C and B.
- (1) The current at point A (increases) (decreases) (remains unchanged).
 - (2) The current at point D (increases) (decreases) (remains unchanged).
 - (3) The potential difference between points A and C (increases) (decreases) (remains unchanged).
 - (4) The potential difference between points F and G (increases) (decreases) (remains unchanged).
- (c) Return to the initial condition shown in the figure. A wire is connected from point E to point C.
- (1) The current at point F (increases) (decreases) (remains unchanged).
 - (2) The potential difference between points C and D (increases) (decreases) (remains unchanged).
 - (3) The current at point H (increases) (decreases) (remains unchanged).

NOTE: As indicated in Section 7.9, many students, including those in engineering-physics courses, hold beliefs about “downstream” effects in electric circuits. The following problem is a way of confronting such beliefs and helping the student rectify them by making a mistake and revising the thinking.

3 Consider the circuit shown in the figure in Question 1. Suppose the connection is broken at point 4, and a resistor is introduced in series with bulbs B and C. What will happen to the brightness of each of the two bulbs B and C? Explain your reasoning briefly.

NOTE: The following may seem to be a trivial question, but it is not. Most students have very substantial difficulty with it. This forms an excellent preparation for discussion of the Wheatstone bridge. It can readily be extended to the case in which the resistances are not identical.

4 The circuit shown in the adjacent figure consists of identical bulbs and a battery with negligible internal resistance. Suppose a wire is connected between points 1 and 2 in the circuit. What happens in this wire? What happens to the brightness of each bulb? What happens to the current drawn from the battery? What happens to the potential difference across each bulb? Explain your reasoning briefly in each case.



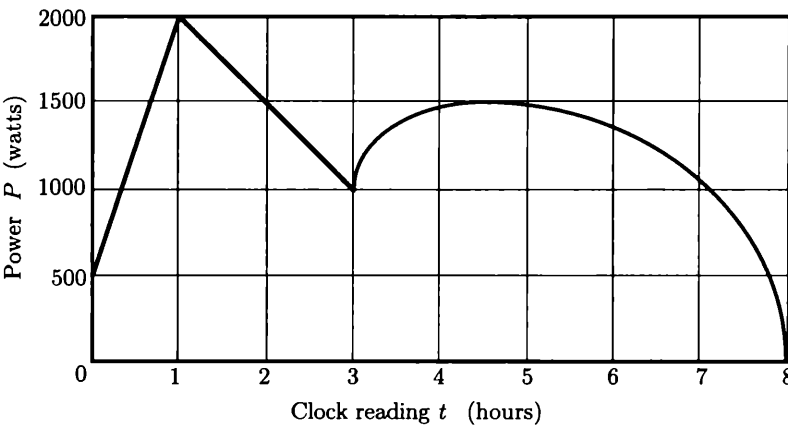
NOTE: The following question affords the opportunity to spiral back to earlier thinking, reasoning, and concept formation. The question may sound trivial to the instructor, but many students, even among those in engineering-physics courses, are astonished to recognize the household meter as a device that measures area under a curve and also to see the connection to apparently unrelated exercises dealt with earlier in the course.

5 The household electric meter is essentially an integrating device—one that measures the area under a curve. It measures the instantaneous power $P(t) = (\Delta V)I(t)$ supplied to the household circuit and calculates, through the integration, the total energy supplied over a period of time.

Shown in the following figure is a hypothetical graph of $P(t)$ versus clock reading t . The potential difference ΔV is kept constant at 120 V.

- What would have been the appearance of the $I(t)$ versus t graph?
- From the figure, obtain a numerical value for the total energy supplied to the household circuit (i.e., the reading in kilowatt-hours that would have been registered on the meter) over the eight hours shown.
- Compare what you have just done in this problem (and the reasoning involved) with an exercise in which you calculate the displacement of a body in rectilinear motion when you have been given a graph of velocity versus clock reading.
- Sketch a graph in which the force acting on an object varies with time, and indicate how you would calculate the impulse delivered by the force over a

specified time interval. Compare this calculation with what you have done in parts (b) and (c).



Chapter 8

Electromagnetism

8.1 INTRODUCTION

It was pointed out in Section 6.2 that many students have not developed, in prior experience, sufficient familiarity with electric and magnetic phenomena to distinguish operationally between them. If care is taken to form the distinction at the earliest introductory levels, as recommended in Chapter 6, understanding of electromagnetism is greatly facilitated. If the operational distinctions have not been clearly formed, however, the electromagnetic phenomena become a source of profound confusion. The recourse becomes, as usual, desperate memorization without comprehension. Among students subject to such confusion, one finds expectations such as the repulsion of positive electric charges by north magnetic poles, attraction or repulsion of stationary charges by the magnetic field around a current carrying wire, and a variety of other nonexistent effects.

Our own knowledge of the phenomena comes not from deduction from abstract principles but from observing the actual effects, discerning the systematic order and relations that are maintained, and memorizing the facts observed. The only way to help students acquire this knowledge is to provide ample opportunity for the visual and kinesthetic experiences that facilitate the necessary memorization of the observed facts. Among the chief sources of failure of instruction in this area of physics are the failure to pay attention to the most basic operational underpinnings and in the failure to allow time for sufficient concrete experience to embed the phenomena in the memory.

Major emphasis in this chapter is placed on ways in which to cultivate experience, and connected thought about this experience, so as to generate the grasp of phenomenology that eludes so many students over the entire spectrum of introductory physics. In addition, it is argued that students should be helped to acquire some understanding of the motivations behind the invention of field theory in the 19th century and the attendant departure from the Newtonian action at a distance view.

8.2 OERSTED'S EXPERIMENT

Although the physical circumstances making up Oersted's experiment are very simple, the basic phenomena are completely unfamiliar to most beginning students, and these phenomena remain very tenuous if viewed only as rapid demonstrations in a lecture room. Fortunately, the magnetic effect around a current carrying wire can be easily studied in simple experiments conducted at home. One needs only a small compass and a circuit consisting of flashlight batteries, bulbs, and wire. Such "take-home" assignments can be readily coupled with the batteries-and-bulbs homework suggested in Section 7.5, and they greatly enrich study that otherwise consists almost exclusively of pushing numbers into end-of-chapter problems. Opportunities for studying real phenomena in homework with simple apparatus are not very plentiful, and there is substantial educational profit in capitalizing on them when possible.

Very few students will conduct a meaningful investigation without guidance, however. The homework assignment should guide them Socratically into: (1) Investigating the compass deflection both above and below the current carrying wire; (2) investigating the effect of reversing the connection to the battery terminals; (3) ascertaining the pattern of the effect all the way around the wire—not just above and below; (4) qualitatively noting the effect of changing the distance between the wire and the compass needle; (5) qualitatively studying the strength of the effect on the compass needle (held at fixed distance from the wire) when additional bulbs are inserted in the circuit either in series or in parallel with an initial single bulb; (6) studying the effect of introducing an additional battery in series with the first; (7) forming, from synthesis of the observations, the right-hand-rule mnemonic for the direction of magnetic field around the current carrying wire.

Out of this simple experience, it is easy to generate understanding of the tangent galvanometer—historically, the first device for comparing the strength of electric currents.

Finally, one can lead the students to study—and thus assimilate a genuine understanding of—the superposition of magnetic effects around current carrying wires. Superposition is rarely discussed explicitly in the textbooks; it is asserted or taken for granted without even a hint that experimental investigation and substantiation are required. Fig. 8.2.1 illustrates how simple this investigation can be made, using the compass at fixed distance from the wire as a crude tangent galvanometer.

Observing the same intensity of effect on the compass needle at positions A, D, and E reinforces the concept of the continuity of electric current all the way around the circuit (i.e., the model initially formed from the batteries-and-bulbs observations outlined in Chapter 7.) The exercise on superposition (comparing the effects at positions A, B, and C) is necessary in its own right, but it is rarely outlined to the students. One can speculate about superposition and hope that nature is simple and linear in this context, but speculation is

not fact. One must appeal to experiment for ultimate sanction. Validation of simple superposition then provides the logical basis for using the right-hand rule to predict the direction of magnetic field around various configurations of the current carrying wire (e.g., coils and solenoids).

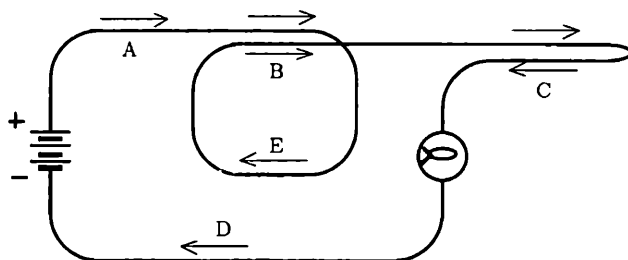


Figure 8.2.1 Superposition of electromagnetic effect: A battery-and-bulb circuit with a long wire bent and laid out in the form shown. At positions A, D, and E, we have a single wire carrying the same current. At position B, we have the effect of two wires carrying current in the same direction. At position C, the two wires carry current in opposite directions. A tangent galvanometer, exhibiting a given torque at positions A, D, and E, exhibits twice the torque at position B and zero torque at position C.

The story of Oersted's serendipitous discovery, in the course of a lecture demonstration, of the effect that now bears his name has become such a commonplace in physics instruction that one tends to lose sight of the fact that deliberate, albeit unsuccessful, search for a connection between electricity and magnetism had a long prior history. Franklin had tried to magnetize a needle by electrical discharge. Whittaker reports in the *History of the Theories of Aether and Electricity* that "In 1774 the Electrical Academy of Bavaria proposed the question 'Is there a real and physical analogy between electric and magnetic forces?' as the subject of a prize." In 1805 two French investigators attempted to determine whether a freely suspended voltaic pile orients itself in any fixed direction relative to the earth. Proponents of Naturphilosophie in the early years of the 19th century hoped to find interconnection among all of the "forces" of nature. Oersted himself was deliberately investigating such connections before his own discovery. It is conducive to insight and learning if students are invited to speculate on these matters prior to abrupt introduction of electromagnetism. Few students have had the intellectual experience of engaging in informed speculation in which one must carefully discriminate speculation from observed fact.

It should be noted that Oersted's experiment introduces an interaction that is profoundly different from interactions students have encountered up to this time. In all the interactions previously encountered (contact, gravitational, electrical, even frictional), the two forces of interaction are colinear and lie along the line connecting the interacting particles. In the Oersted experiment,

the forces on the poles of the compass needle are orthogonal to the radial line from the current carrying wire to the point at which forces are being observed. The interaction is that of torques rather than colinear forces. Very few textbooks or presentations call this fact to the attention of the students; yet this is a deeply significant extension of previous experience.

Finally, students should be led to articulate the question as to whether moving electrostatic charges would produce the same effects as current in a wire. That this is not a trivial question is indicated by the discussion that attended it during the 19th century. Although most physicists anticipated an affirmative answer, and although Faraday firmly asserted in 1838, that "if a ball be electrified positively in the middle of a room and be then moved in any direction, effects will be produced as if a current in the same direction had existed," the effect remained to be confirmed experimentally. This was, of course, not an easy task because of the weakness of the effect—eloquent testimony to the relatively enormous amounts of charge transported in electrical conduction. The direct test was finally performed by H. A. Rowland in 1875 while he was spending a year in Helmholtz's laboratory in Berlin before returning to the United States to assume the professorship of physics at the newly founded Johns Hopkins University. Rowland charged the periphery of a rapidly rotating nonconducting disk and showed that the magnetic field around the rim was essentially the same as that around the current carrying conductor.

An effective homework question evolves from a description of some of Oersted's original experiments:

Among his many experiments and observations, Oersted reports the following: (a) "The kind of metal forming the conductor does not alter the effects, except, perhaps as regards their intensity. We have employed with equal success wires of platinum, gold, silver, copper, iron, bands of lead and tin, and a mass of mercury." (b) The effects on the compass needle remain virtually unaffected when rock, wood, glass are placed between the wire and the needle, and when the needle is encased in a copper box full of water. (c) "Needles of copper, glass, and resin, suspended like the magnetic needle, are not affected by the current carrying wire."

Interpret Oersted's experiments: Why did he perform them? What conclusions are to be drawn from the reported observations?

Students are rarely afforded the opportunity to confront questions dealing with the point and interpretation of qualitative observations such as those described by Oersted. Here is a valuable learning experience to be added to that afforded by the conventional problems.

8.3 FORCES BETWEEN MAGNETS AND CURRENT CARRYING CONDUCTORS

In the usual performance of Oersted's experiment, the use of stiff or massive wires and a small, sensitive compass focusses all attention on the compass and tends to conceal the fact that forces exerted on the compass are accompanied by forces exerted on the conductor. Oersted himself was among the first to point out that: "As a body cannot put another in motion without being moved in its turn, when it possesses the requisite mobility, it is easy to foresee that the current carrying wire must be moved by the magnet." Although this was "easy" for Oersted to foresee, it is the very rare student (only a budding research physicist) who thinks of the possibility without prompting.

Socratic prompting can be very effective in this context, especially if very little is given away. Students derive considerable satisfaction from bridging a gap even if the latter is small, and such satisfaction should not be deprecated. Some students are even able to predict the direction of force on the wire starting from the Oersted effect.

Lecture demonstrations of the force on the current carrying conductor in the field of a magnet are almost universally performed, but, in many instances, their effectiveness is limited or diluted for a number of reasons. Sometimes the link to the Oersted experiment is not made clear. In other instances, the demonstrations are performed far too rapidly, and students do not have time to register the multiplicity of directions involved. The greatest lack, however, tends to be the lack of opportunity for the individual student actually to feel the tug on the wire. Despite the fact that they have seen the wire jump and accelerate when the switch is closed, students show astonishment, bordering on incredulity, when they hold the wire and feel the pull. Making this opportunity available to all students is most desirable if at all possible. The usual mnemonic (the cross product, or second right-hand, rule) is most effectively registered if students have the opportunity to put it together out of their own observations and sensations close up rather than from the remoteness of the lecture demonstration.

A useful and impressive demonstration that does not seem to be widely performed is that of showing the interaction between a magnet and a current carrying electrolyte, that is, an interaction with electric current not confined to a metallic conductor—one in which the current is *not* electron current, but positive and negative ions are moving in opposite directions. The force on the current carrying conductor produces a flow of the electrolyte in the direction predicted by the cross product rule for the case of conventional positive current. The experiment is quite simple, and a basic arrangement is sketched in Fig. 8.3.1. The effect is easily displayed on an overhead projector.

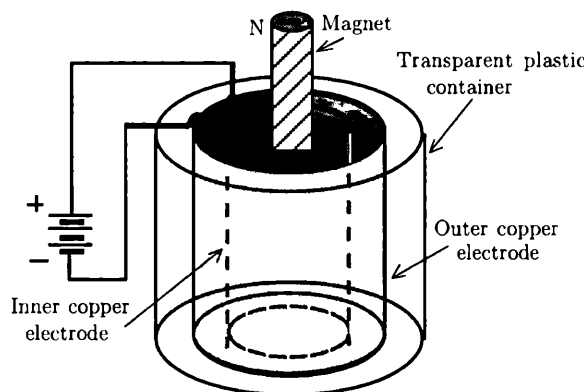


Figure 8.3.1 Interaction of magnet with current carrying electrolyte: Copper sulfate solution is placed in a plastic container between two cylindrical copper electrodes, which are connected to a battery or other power supply. When the end of a rod or bar magnet is inserted at the center of the system (within the inner electrode), the liquid is observed to circulate in the annular space between the electrodes. The circulation can be made visible by sprinkling wood filings or some floating powder on the surface. The direction of circulation is consistent with the right-hand rule devised in connection with the force on current carrying wires.

8.4 AMPERE'S EXPERIMENT

Another consequence of Oersted's discovery is the expectation that two current carrying conductors might exert forces on each other since the magnetic effect produced by one would interact with the current of the other, and vice versa. It should be noted, however, that this is merely a plausible argument and does not constitute a reliable inference. As was pointed out by Arago, a magnet exerts a force on each of two unmagnetized pieces of iron, but the pieces of iron, when separated from the magnet, exert no force on each other. The situation with the current carrying conductors could be analogous: Namely that the presence of a magnet is essential and that, without the magnet, the current carrying conductors are as neutral as the pieces of iron. Experimental verification is necessary, and it was supplied by Ampere within a week after the news of Oersted's discovery reached France.

Ampere reported to the French Academy the results of the experiment now universally reproduced in lecture demonstrations: The attraction between parallel wires carrying current in the same direction and the repulsion with currents in opposite directions. There is a profound difference, however, between the tone of Ampere's report and that of most textbook and lecture presentations of the same phenomenon. Ampere was reporting the discovery of a new phenomenon, and, since batteries and electricity were involved, he had to provide convincing evidence that the interaction was not simply electrostatic

but justified the designation “electromagnetic,” which he introduced. It is this aspect of the phenomenology that is largely ignored in modern presentations: The effect is immediately asserted to be “electromagnetic” without further inquiry and at substantial cost in the opportunity for students to think and learn.

It is far better pedagogy to exhibit the interaction and then to lead students into raising the question as to whether the effect is new and different or simply electrostatic. Once the question is raised, students must confront the actual observations in full detail, suggest additional experiments, and outline in operational terms the differences between what is actually observed and what would be expected of ordinary electrostatic interaction. This engages them in purely phenomenological thinking, without formulas and number grinding, and forces them to review and reconsider earlier experiences. As has been argued repeatedly, such spiralling back in richer and more sophisticated context is at the leading edge of genuine learning and understanding. Homework can be greatly enriched by consideration of such questions since they induce thinking about all aspects of the related (and unrelated) phenomena—not just dealing with the restricted end results.

Ampere regarded the question as far from trivial and argued his case in the following way:

These attractions and repulsions between electric currents differ fundamentally from the effects produced by electricity in repose. First, they cease, as chemical decompositions do, as soon as we break the circuit [i.e., he argues that the effect is dynamic rather than static]. Second, in ordinary electric attractions and repulsions, opposite charges attract, and like charges repel; in the attractions and repulsions of electric currents, we have precisely the contrary. It is when two conducting wires are placed parallel in such a way that their ends of the same sign are next to each other that there is attraction, and there is repulsion when the ends of the same sign are as far apart as possible. Third, in the case of attraction, when it is sufficiently strong to bring the movable conductor in contact with the fixed conductor, they remain attached to each other like two magnets and do not separate after a while, as happens when two conducting bodies, oppositely electrified, come to touch [we tend to perform our demonstrations with insulated wires, but Ampere was using bare conductors that could have transferred charge between each other.]

One might add to this list the observation that three parallel wires, carrying current in the same direction, all attract each other. This is an effect that cannot possibly occur electrostatically. (It is interesting that Ampere was aware of this effect but does not adduce it along with the other three listed in the preceding paragraph.)

The phenomenological arguments that the interaction between parallel current carrying wires cannot be electrostatic lend themselves to formation of a very effective homework problem, especially if Socratic guidance is provided without giving the whole story away.

8.5 MNEMONICS AND THE COMPUTER

Virtually all courses and textbook presentations introduce some form of the simple mnemonics that help one keep track of electromagnetic phenomena: The right-hand rules for the Oersted effect and for the Lorentz force on moving charges (and the force on a current carrying conductor in the field of a magnet) and some form of left-hand rule for Faraday's law of electromagnetic induction. Many students, however, fail to master these rules and fail to use them correctly either in concrete or in pencil-and-paper situations.

This failure is clearly not one of logical reasoning or of concept formation; it is simply a matter of practice and memory. Textbooks do not, in general, provide a sufficient number of exercises for application of the rules, and this form of knowledge is not frequently tested. Students do not generate their own exercises and practice—even though they are perfectly capable of doing so if guided into the process and if they know they will be tested on the material.

What is needed here is a certain amount of drill, and the term “drill” is not being used pejoratively. It is only through drill that such knowledge can be fixed in the memory. Drill should take the form of applying the rules in different orientations of the motions and field directions in space. Drill should also lead the student to confront all possible variations of the rules: Given the direction of current in a straight wire (or in a coil), what must be the direction of the magnetic field at various indicated points? Given the direction of magnetic field at some point in the neighborhood of a straight wire or of a coil, what must be the direction of conventional current in the conductor? Given the direction of motion of a charged particle and the direction of the Lorentz force, what must be the direction of the magnetic field? (And the other two combinations of knowns and unknown.)

Students should also have the opportunity to sketch trajectories of moving charged particles in magnetic fields. They frequently say what sound like the correct words in describing such motions, but they proceed to draw trajectories that are physically impossible and fail to recognize what is wrong with their drawings.

It is in such areas of drill that the computer can provide very effective assistance to learning. Interactive dialogues with relatively simple graphics can provide students with as many exercises as are needed to master the rules, and they can help save a great deal of time. This is one of the potentially cost-effective areas of computer-based instruction.

8.6 FARADAY’S LAW IN A MULTIPLY CONNECTED REGION

Figs. 8.6.1 and 8.6.2 show a simple and very fundamental physical situation that is rarely, if ever, considered in introductory physics. Yet the phenomena arising in this system are deeply significant and greatly enhance a learner’s understanding of essential aspects of electromagnetism.

A long solenoid (Fig. 8.6.1), axis perpendicular to the plane of the paper, carries a varying current. Since the magnetic flux through any conducting loop surrounding the solenoid changes with time, an emf is induced (counterclockwise) in the conducting loop in accordance with Faraday’s law. Flashlight bulbs in this loop will light; potential drops will develop across resistors.

Figure 8.6.1 A long solenoid, axis perpendicular to the plane of the paper, carries a varying current. An emf and a resulting current are induced (counterclockwise) in the conducting loop surrounding the solenoid. Flashlight bulbs in this loop will light; potential drops will develop across resistors.

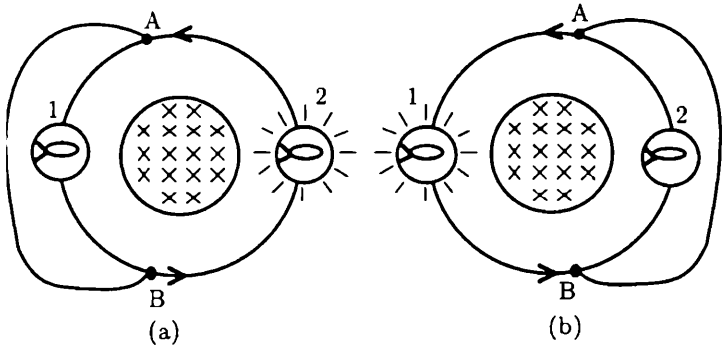
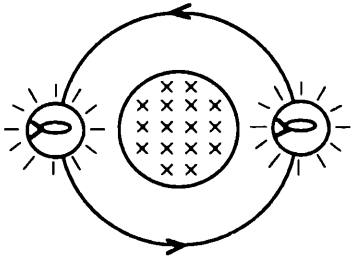


Figure 8.6.2 When points A and B are shorted by a wire as in (a), bulb 1 goes out while bulb 2 remains lighted and burns more brightly than before. When the same points are shorted as in (b), bulb 2 goes out while bulb 1 remains lighted and burns more brightly than before.

Suppose that a wire is now introduced in such a way as to short circuit points A and B in the loop (Fig. 8.6.2). Students’ prior experience with simple batteries-and-bulbs circuits has led to the expectation that short circuited bulbs will always go out. The situation now being considered is, however, very different because the region outside the solenoid is multiply, rather than simply, connected. Only one bulb goes out, and the other burns more brightly, depending on whether the shorting is effected as in Fig. 8.6.2(a) or as in Fig.

8.6.2(b). Thus, in this situation, naive expectation is contradicted, and one must think through what happens in terms of the change of flux through the various conducting loops in the system.

Romer (1982) discusses this problem in complete and sophisticated detail in terms of Maxwell's equations applied to various regions in the system and in terms of interpretation of the readings on voltmeters connected between points A and B. Readers interested in pursuing, in full detail, the issues raised here would do well to refer to Romer's lucid paper. The presentation given above is merely a qualitative, first stage, introduction to a significant problem. This introduction, even if incompletely rigorous, extends the range of student understanding by broadening and enriching the context.

8.7 FARADAY'S CRITICISM OF ACTION AT A DISTANCE

Among Faraday's experimental achievements were the recognition of several new chemical compounds; work on the liquefaction of gases; discovery of the laws of electrolysis; discovery of the phenomena of electromagnetic induction, diamagnetism, and the rotation of the plane of polarization of a beam of light by a magnetic field. Faraday's great influence on subsequent scientific thought stemmed, however, not only from the importance of his experimental discoveries but also from the influence of his theoretical speculations. His investigations led him to reject Newtonian action at a distance in connection with electric and magnetic interactions and to assign a role to lines of force in the intervening medium. In doing so, he laid the basis for the modern concept of "field," which was elaborated mathematically by Maxwell and which has become one of the most firmly established elements in all of theoretical physics.

Prior to the success and acceptance of the Newtonian theory of gravitation, the concept of action at a distance was anathema to most natural philosophers. In their revulsion against superstition and against the assignment of occult virtues and properties to inanimate substances, they rejected the thesis that a material object could exert an effect in a place where it was not. They tried to visualize clear-cut mechanisms for all interactions and, where an intervening medium was not directly apparent to the senses, they invented effluvia and continuous or corpuscular ethers through which effects could be propagated by contiguous action of layer on layer or particle on particle. The word "attraction," in certain contexts, for example, implied action at a distance and was rejected as yielding to occult beliefs. When Huygens was informed of the imminent publication of Newton's *Principia*, it is said that he expressed the hope that it would not be based on "attractions," as rumors had that it would be. Huygens and Leibniz never accepted Newtonian theory on this account. The opposition, of course, eventually died out because of the repeated successes of the theory, and action at a distance (the applicability of Newton's third law instant by instant throughout a gravitational interaction) came to be accepted uncritically. Max Born remarks in this connection that,

" . . . after Newton's theory . . . had been established, the idea of a force acting at a distance became a habit of thought. For it is indeed nothing more than a habit when an idea impresses itself so strongly on minds that it is used as an ultimate principle of explanation."

Newton himself had been careful to avoid commitment concerning models of gravitational interaction—"Hypotheses non fingo." He was clearly aware that he had discovered a mathematical formulation that correctly represented the facts of natural phenomena but that might be subject to a variety of interpretations.¹ Newton himself probably did not believe literally in action at a distance. He engaged in much speculation about ethers through which gravitational and other effects might be propagated, and there exist among his writings chapters that were originally intended for the *Principia* but that he withheld, possibly because he felt their speculative nature to be out of harmony with the Euclidean tone of the rest of the work.

Ampere constructed his mathematical theory of electromagnetic interaction between current carrying elements of wire in the action at a distance framework. Although Faraday in the *Electrical Researches* repeatedly expresses admiration for the "beautiful work" of Ampere and other members of the French school, his own fundamental outlook was quite different. Lacking formal training, especially in mathematics, he was never proficient in the abstract language of analysis (calculus). He tended to interpret his observations and formulate his concepts almost entirely in geometrical and physical terms. It is possibly because of this orientation that the geometrical patterns formed by iron filings impressed him so deeply. Faraday began to refer to "lines or curves of magnetic force" (a term that had already been used in connection with magnetism at least as early as the beginning of the 17th century): "By magnetic curves, I mean lines of magnetic forces, however modified by the juxtaposition of poles, which could be depicted by iron filings; or those to which a very small magnetic needle would form a tangent." Later he extended the idea to include lines of electric force.

Faraday's observations and experiments convinced him that the totality of electric and magnetic phenomena could not be explained in terms of action at a distance between particles, and that the intervening space must somehow be involved. Among the pieces of evidence that seem to have weighed most heavily with him are the following:

- 1 The lines of force are curved in space and are not simply straight lines connecting interacting magnetic poles or charged particles. Faraday could

¹Situations of this kind have arisen from time to time in the history of science. In 1822, for example, Fourier published a complete and elaborate mathematical theory of the conduction of heat in solids (this was the work in which the Fourier series was presented.) Fourier thought and wrote in terms of the caloric theory of heat, which was soon to be rejected, but his differential equation correctly represents the conduction of heat regardless of what model we hold as to the nature of heat itself. His theory was therefore in complete agreement with the observed temperature distributions.

not conceive the curvature of lines of force in terms of action at a distance, especially in the case of lines around the current carrying wire. He felt that the curvature either indicated a physical existence of the lines themselves, or else stemmed from a state or condition of an intervening ethereal medium.

2 When a slab of dielectric material (Faraday used discs of sulfur and shellac) was inserted between the plates of a charged capacitor, the quantity of charge on the plates was found to be different from what it had been when only air intervened. To Faraday, this experiment indicated that interaction between electrical charges on the plates was not the exclusive factor and that the intervening medium also played an important role. He was also aware of similar effects in magnetism.

3 In connection with induction of electric current in a wire loop when a portion of the wire is moved so as to cut lines of magnetic force (motional emf), he says, "The mere fact of motion cannot have produced this current: there must have been a state or condition around the magnet and sustained by it, within the range of which the wire was placed."

The fact that one might perhaps take issue with some of these views and rationalize the observations in some other way is not the point. These were the factors that influenced his phenomenological thinking.

The most crucial question, that to which Faraday returned repeatedly, was whether or not finite time intervals were required for the propagation of changes in electric and magnetic effects: When current was abruptly changed in a wire, did a time interval elapse before the magnetic lines of force changed some distance away? When quantity of charge was abruptly changed on an electrically charged body, did a time interval elapse before a change in the force exerted on a test particle some distance away? If a charged particle were abruptly displaced, would a time interval elapse before a change in the force exerted on a stationary test particle some distance away?

An affirmative answer to these questions would have lent the strongest possible support to Faraday's views concerning the dubious status of action at a distance models since it would have implied a propagation of action, perhaps in wavelike fashion, through the intervening space. (Although Maxwell's theory, developed over the decade 1856-65, did answer these questions in the affirmative, direct experimental verification of the theory did not come until Hertz's famous experiment in 1887, twenty years after Faraday's death.)

By the time Faraday was pursuing this line of thought, Young, Fresnel, and others had shown that light behaves like a transverse wave. It was far from clear just what was waving, but many physicists suspected and searched for a connection between light and electricity and magnetism. Faraday himself discovered the first direct physical interaction: The rotation of the plane of polarization of light when propagating in a direction parallel to magnetic lines of force. He entertained the idea that light might itself be the transverse vibrations of lines of force. He speculated on whether the velocity of propaga-

tion of magnetic effects might be of the same order as the velocity of light, and these thoughts provided further motivation for rejecting action at a distance and tentatively assigning a fundamental role in electromagnetic phenomena to lines of force or to an ethereal medium.

In his biography of Faraday, Tyndall writes,

During the evening of his life, he brooded on magnetic media and lines of force; and the great object of the last investigation he ever undertook was the decision of the question whether magnetic force requires time for its propagation. How he proposed to attack this subject we shall never know. But he left some beautiful apparatus behind; delicate wheels and pinions, and associated mirrors, which were to have been employed in the investigation.

In retrospect, we know that the technology of the time was not ready for such an experiment.

Readers will note that, in earlier chapters, I have repeatedly advocated directing the attention of students to questions concerning Newton's third law, action at a distance, and time intervals associated with interactions. The point, of course, is to prepare students to understand and appreciate this bit of intellectual history—to comprehend Faraday's questions and the motivation for invention of field theory.

8.8 INFANCY OF THE "FIELD" CONCEPT

Faraday, in his writings, carefully separated speculations regarding ethers and lines of force from factual reports of the results of his experiments, and he hedged these speculations almost apologetically:

It is not to be supposed for a moment that speculations of this kind are useless or necessarily hurtful in natural philosophy. They should ever be held as doubtful and liable to error or to change, but they are wonderful aids in the hands of the experimentalist and mathematician; for not only are they useful in rendering the vague idea more clear for the time, giving it something like a definite shape, that it may be submitted to experiment and calculation; but they lead on, by deduction and correction, to the discovery of new phenomena, and so cause an increase and advance of real physical truth, which unlike the hypothesis that led to it, becomes fundamental knowledge not subject to change.²

²This is an especially valuable paragraph to have the students read, think about, and discuss. One can hardly find a more cogent or more lucid description of the power and utility of a heuristic device in scientific thought. On the other hand, the end of the paragraph uses phrases that very few scientists would use today. Faraday was by no means the only scientist of his time to believe that science produces "real physical truth" and "knowledge not subject

Nevertheless, it is evident in much of Faraday’s subsequent work that lines of force meant more to him than just a heuristic device. He used this idea so much and with such success that he clearly came to believe in “physical” lines of force. In this conception the lines filled all space and had distinct properties and modes of behavior. They acted like rubber bands that were under tension longitudinally and that repelled each other laterally. William Thomson (Lord Kelvin), in papers published in 1847 and 1854, called attention to mathematical analogies that exist between theories of fluid flow, heat flow, and elasticity on the one hand and electrostatics and magnetism, as described by lines of force, on the other. Faraday, taking mathematical analogy to other physical phenomena as evidence of physical reality, felt his view of lines of force to be strongly supported.

James Clerk Maxwell, then a young Fellow at Trinity College, Cambridge, was deeply impressed both with Faraday’s conception of lines of force and with Thomson’s revealing mathematical analogies. Gifted with great mathematical talent and with intuitive physical sense on a par with Faraday’s, he embarked on an attempt to synthesize into one unified theory all the known phenomena of electricity and magnetism. In his first two papers on this subject, published in 1856 and 1861, he developed an elaborate fluid model of Faraday’s lines of force:

By referring everything to the purely geometrical idea of motion of an imaginary fluid, I hope to attain generality and precision, and to avoid the dangers arising from a premature theory professing to explain the cause of phenomena. If the results of mere speculation which I have collected are found to be of any use to experimental philosophers in arranging and interpreting their results, they have served their purpose, and a mature theory, in which physical facts will be physically explained, will be formed by those who, by interrogating Nature herself, can obtain the only true solution of the questions which the mathematical theory suggests.

Note the similarity of this remark to the first part of the statement by Faraday quoted above. In his first paper, Maxwell used an elaborate concrete model involving fluid cells, vortices, and “idler wheels.” A diagram of such a system

to change.” With only a few dissenting voices influenced by the positivistic movement in philosophy (Mach, Ostwald, Duhem, for example), many 19th century scientists would have reflected similar attitudes. Their confidence is quite understandable. So wide in scope, so convincing were the successful applications of Newtonian mechanics and the new theories of thermodynamics and electromagnetism, that they indeed seemed to have led to “knowledge not subject to change.” The warnings of the skeptics went largely unheeded. When one reaches the point of discussing the early 20th century revolutions in relativity and in atomic and quantum physics, it is very effective to lead students back to this paragraph and to contrast our modern view of the provisional nature of our knowledge with Faraday’s confident assertion.

is given in the paper. In these papers, Maxwell also began to use the terms “electric field” and “magnetic field” in the essentially modern sense. In 1865 he published his final version of the theory, explicitly eschewing action at a distance:

I have preferred to seek an explanation [of electric and magnetic phenomena] by supposing them to be produced by actions which go on in the surrounding medium as well as in the excited bodies, and endeavoring to explain the action between distant bodies without assuming the existence of forces capable of acting directly at sensible distances.

The theory I propose may therefore be called a theory of the ‘Electromagnetic Field’ because it has to do with the space in the neighborhood of the electric and magnetic bodies, and it may be called a ‘Dynamical’ theory because it assumes that in that space there is matter in motion, by which the observed electromagnetic phenomena are produced . . . [The space] may be filled with any kind of matter, or we may endeavor to render it empty of all gross matter, as in the case of Geissler [electrical discharge] tubes and other so-called vacua.

In this paper (as well as in the subsequent *Treatise on Electricity and Magnetism* published in 1873), the elaborate fluid model of cells and vortices has disappeared. There remain only the mathematical equations and the concept of “field” as a condition or state of an ethereal medium. General acceptance of Maxwell’s theory toward the end of the 19th century marked the transition from an era dominated by action at a distance philosophy to the present era of field theories in which momentum, energy, and other conserved quantities are propagated through the “field.”

Students who go on to more advanced studies in physics and engineering should certainly be helped to acquire some of this phenomenological background so as to better understand and appreciate the real point and purpose behind the introduction of the “field” concept. Such background makes the introduction and the consequences of Maxwell’s equations much more understandable and intelligible. However, even students who never go on to more advanced study and will never see or use Maxwell’s equations can benefit from this qualitative exposure. Given the build-up and tie-ins urged in earlier chapters, they can acquire at least a qualitative insight into the revolutionary shift in point of view that accompanied the invention of field theory and have some comprehension of what questions and points of view motivate the modern search for evidence of gravity waves and how it comes about that the field theoretic point of view permeates all of quantum mechanics and particle physics.

8.9 LABORATORY MEASUREMENT OF A VALUE OF B

One of the few measurements worth making just for the sake of the measurement itself is that of the strength of a B-field. The reason for this is that very few students develop confidence in the meaning of B directly from text or lecture presentations. If questioned, many actually reveal doubts about the “reality” of such numbers. Such doubts and reservations are markedly reduced by the concrete experience of making a direct measurement of the force acting on a current carrying wire in a magnetic field and calculating the magnitude of B from F/IL . (Flip coil or Hall effect devices do not have the same impact since the connection to B is more abstract and more remote.)

Most apparatus companies offer a device for determining B by “weighing” the force on a known length of current carrying wire. *PSSC Physics* used to exploit a simple homemade device that was quite effective. A simple, flexible setup offers the advantage of allowing students to explore the variation of the force with the angle between the field direction and the wire—something worth doing in addition to making the force measurement when the two directions are orthogonal.

After making such observations themselves, students hold a markedly different attitude toward the meaning of B than they hold in the absence of the concrete experience.

Chapter 9

Waves and Light

9.1 INTRODUCTION

The teaching of wave phenomena in introductory physics rightly concentrates on kinematic aspects and leaves most of the dynamics to later, more advanced consideration. One of the pedagogically best and soundest treatments is still that of the *PSSC Physics* textbook in its six editions. Fortunately, many other books have drawn heavily on the *PSSC* treatment, especially the fine photographs that have been made widely available. When combined with demonstrations, laboratory experience with ripple tanks, and the collection of excellent film loops and videotapes showing reflection, refraction, and interference of ripples, such text material is quite effective in generating understanding of wave behavior and the distinction between wave and particle motion. Excessively rapid coverage of this material, however, frequently negates the potential effectiveness, especially when direct laboratory experience with strings, slinkies, and ripple tanks is not made available, and when accompanying Socratic questioning is not provided.

Concrete experience is still an essential factor in cultivating understanding of the phenomena and grasp of the extensive vocabulary that is generated. Furthermore, this experience must be guided by phenomenological questioning of a kind that is missing in many textbook and lecture presentations. The following sections contain some examples of what might be done to fill a few of the more serious remaining gaps. Examples are also given of insights that enrich the context and lead students to become aware of deep connections among seemingly disparate phenomena.

9.2 DISTINGUISHING BETWEEN PARTICLE AND PROPAGATION VELOCITIES

Although, in the case of the transverse wave on a string, the distinction between particle velocities in the medium and propagation velocity of the distur-

bance seems quite obvious visually, some students still exhibit residual confusion despite demonstrations they may have seen. They readily admit that the velocities are orthogonal to each other, but they fail to discern that the magnitudes are quite different. Some of this confusion is associated with failure to perceive that the particle velocities vary in both magnitude and direction and do not possess a single unique value (as does the propagation velocity). Some confusion stems from the fact that the maximum particle velocities increase and decrease together with increases and decreases in propagation velocity as the tension in the string is varied.

Lecture demonstrations on these matters tend to go by too rapidly for slower individuals, and such students should be helped to confront these aspects in observations of their own, preferably as part of home experiments with strings or ropes. Such homework, however, needs to be structured so as to guide students into genuine observation rather than vacuous “playing around.” They should, for example, be led to see that the particle velocities differ from the propagation velocity in magnitude and that the velocity of the particle is zero at maximum deflection. They should sketch the variation in particle velocity through positive and negative pulses traveling in both possible directions. Initially, it is sufficient that such sketches be qualitatively adequate; to require that they be rigorously correct is asking too much at so early a stage.

In the case of longitudinal waves, many more students fail to discriminate between particle and propagation velocities because of the colinearity of the two. Rapidity of coverage is again a frequent obstacle. Most teachers are aware of how helpful the soft coil spring called the “slinky” can be in this context. Students master these ideas if given a reasonable amount of time to handle and observe the slinky and if they are required to sketch the variation of particle velocity in both compression and rarefaction pulses. The insight does not develop, however, unless suitable qualitative questions and problems are provided, and texts, by and large, do not supply the needed guidance. The task devolves on the teacher.

9.3 GRAPHS

As in the case of rectilinear kinematics (see Sections 2.6 and 2.10), the sketching and interpretation of graphs can play a key role in developing the student’s understanding of wave kinematics—as well as understanding of many text presentations that are not otherwise assimilated.

Since most texts concentrate almost entirely on sinusoidal wave forms (which, of course, happen to have a simple analytical representation) and tend to ignore arbitrary pulse shapes (which do not), many students fail to develop a clear distinction between graphs in which the abscissa represents clock reading (i.e., passage of time at a fixed location in the medium) and graphs in which the abscissa represents position along the x -axis at a fixed

clock reading. The simple way to deal with this is to add homework problems and test questions such as the following: (1) Given the “photograph” (y versus x graph) of an asymmetric pulse on a string, sketch the corresponding y versus t graph; (2) given the y versus t graph, sketch the y versus x graph. (It is important that the shape be asymmetric because otherwise the distinction between the two representations is lost.) The corresponding particle velocity graphs should also be sketched. (See Fig. 9.3.1 for an illustration of what is being described.)

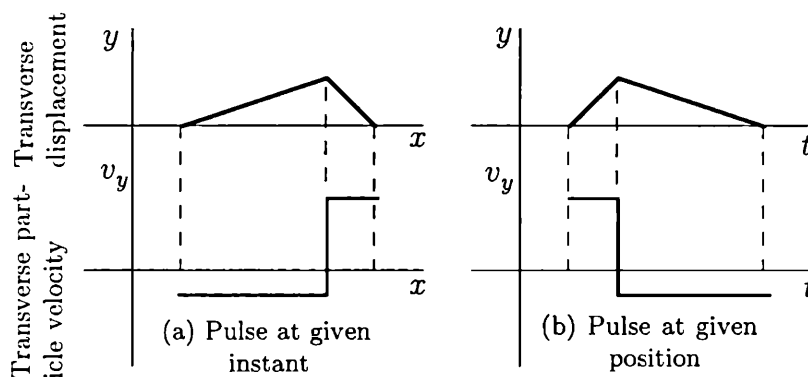


Figure 9.3.1 Corresponding graphs of an asymmetric pulse on a spring:
 (a) As a function of position along the string at a fixed instant of time;
 (b) as a function of clock reading at a fixed position along the string.

Problems of this variety are relatively simple for the case of transverse waves on a string because particle displacements are directly represented by ordinates of the graphs (i.e., the shape of the y versus x graph is the same as the shape seen on the string except for a change in scale). Even so, many students have difficulty with the transformation to the y versus t graph as well as with the velocity graphs, especially if they did not have adequate practice with kinematic graphs earlier in the course.

The difficulties are compounded, however, and affect many more students, when the transition is made to longitudinal waves. A major source of difficulty is now the fact that the ordinate in the initial graphs (before one makes the transition to particle velocity) is some variable such as pressure, or density, or particle displacement from equilibrium position, and the “shape” of the pulse is no longer directly visible as it is on a string.

Students’ understanding of longitudinal waves can be greatly strengthened if they are supplied with qualitative questions that present asymmetric pulse shapes and call for (1) interpretation of the graph by means of a picture showing the corresponding variations in the medium (i.e., closer or wider spacing of coils of the slinky; higher or lower pressure or density indicated by closer or wider spacing of dots representing molecules of gas); (2) transformation of x into t graphs and vice versa.

The sketching of corresponding particle velocity graphs helps register one of the most significant physical distinctions between longitudinal and transverse waves: Although the particle velocity is zero at the point of maximum deflection in the transverse pulse, the particle velocity is at a maximum at the point of maximum compression in a longitudinal pulse.

After such exercises have been performed with asymmetric pulses, they should be performed with the sinusoidal wave trains that become the principal burden of subsequent discussion, but the pulses are important because they emphasize physical aspects that are blurred in the sinusoidal trains. The pulses are also helpful in cultivating better understanding of transverse and longitudinal wave reflections at free and rigid boundaries, a subject dealt with in Section 9.5.

9.4 TRANSVERSE AND LONGITUDINAL PULSE SHAPES

Section 9.3 pointed out the value of invoking asymmetric pulses, in addition to sinusoidal wave trains, in helping students visualize important physical details of wave phenomena through exercises in graphing. Qualitative consideration of the actual generation of simple pulse shapes helps bring out additional physical aspects that are rarely made clear in introductory physics but that play a significant role in developing better understanding of the phenomena.

In the case of transverse waves on a string, either a purely positive or a purely negative pulse can be generated by deflecting the end of the string in either the positive or negative direction and bringing the end back to the zero position. To generate a pulse with both a positive and a negative phase, it is necessary to swing the end of the string to both positive and negative positions before returning to zero.

The situation is quite different in the case of longitudinal pulses. If one wishes to generate a pure compression pulse on the slinky, for example, one must create a compression by moving the end of the spring and leaving the end displaced in the direction of compression. If one moves the end back to the zero position, the compression phase is inevitably followed by a rarefaction phase. Similarly, if one wishes to generate a pure rarefaction pulse, one must create the rarefaction by leaving the end permanently displaced in the direction of rarefaction; if the end is returned to zero position, the rarefaction will be followed by a compression phase.¹

In observing students doing laboratory work with wave phenomena, I have seen significant gains in security and confidence when these aspects of pulse generation become part of their basic understanding.

¹These qualitative observations are, of course, directly related to a very basic theorem regarding wave propagation, namely that the net *impulse* carried by a wave must be zero if there is zero final displacement at the point of origin. I do not advocate developing this theorem quantitatively in an introductory course, but the qualitative insights being suggested here provide a firm basis for deeper understanding at later, more advanced levels

One aspect of understanding of acoustic compression and rarefaction pulses deserves special attention. When asked to describe in their own words what is happening in the fluid medium as such pulses are generated and then propagate, many students respond with description at a microscopic (atoms and molecules) rather than at a macroscopic level. Although there is nothing intrinsically wrong with such a description (except for the fact that students rarely, if ever, visualize the chaos of thermal motion superposed on the organized wave behavior), recourse to it usually reveals a deeply seated reluctance to deal with the macroscopic properties of pressure and density. The jargon about atoms and molecules has been picked up in earlier schooling, frequently with distorted and misleading overtones, while the macroscopic properties—especially pressure—are not well understood.

Hydrostatic pressure is, in fact, a subtle and difficult concept. Physical understanding of phenomena involving pressure and pressure variations hinges on an understanding of Pascal's law—the fact that pressure at a point in a fluid is uniform in all directions. Without explicit grasp of this concept, many aspects of what happens in fluid media are imperfectly understood and visualized. With the modern tendency to omit or shortcut the study of fluid phenomena in introductory physics, many students emerge with very weak understanding of the nature of fluid pressure. (Witness the failure of many practicing physicists to recognize that, when an oil-water mixture separates on standing, the pressure changes on the bottom of the container if the container has sloping sides.) The question of understanding of fluid pressure is discussed in more detail in Section 11.3.

9.5 REFLECTION OF PULSES

Reflection of waves at boundaries is a very different physical problem from that of particles bouncing off walls. With waves, one is dealing with a boundary value problem in a dynamical system, and it includes all the subtlety associated with partial reflection and transmission and with absorption, cases of complete reflection being only idealized limits. Simultaneous reflection and transmission is one of the intrinsic properties separating wave and particle behavior in classical physics.

In the introductory course, one cannot proceed to develop the formal mathematical solution of the wave equation at the boundary even for the idealized cases. Visualization of the reflected wave requires something of an ad hoc argument, and the argument tends to be glossed over in many text presentations. As a result, many students are mystified by the approach—that of visualizing a reflected wave propagating out of the “never never land” on the other side of the boundary and having a shape such as to satisfy the boundary condition. They feel this technique to have a touch of black magic and to be something they themselves could never have conceived.

It may not be possible to allay such doubts completely, but it helps to say something explicit and to motivate the approach. The starting point for such motivation is simply the observed fact that reflections do occur and that the direction of propagation is opposite to that of the incident wave. Since, in the ideal case, the wave in one dimension propagates without change in shape, one can visualize the incident wave as having started or come from anywhere along the string or spring or water surface, even from a region into which the medium in question does not actually extend physically. In just the same way, it becomes legitimate to visualize the reflected wave as having been propagating without change in shape from anywhere. This helps justify visualizing the reflected wave as propagating in the region beyond the reflecting boundary (whether there is any medium there or not) and arriving at the boundary in such phase as to maintain the free or rigid boundary condition.

One must confess explicitly that this is a purely kinematic approach that helps “save the appearances” (to use an ancient locution), one that accepts the existence of reflection as an observed fact and serves to satisfy the condition at the boundary. It is not a dynamic approach, and it therefore contains no description of a “mechanism” by which the reflection is generated at the boundary. (Most attempts to provide a mechanical explanation of the form of the reflected wave are misleading or specious, and it is better that they be avoided.) Most students are willing to accept the procedure being adopted when one is frank about its justification and limitations.

Although many textbooks, especially those making use of the excellent *PSSC Physics* illustrations, do give good, clear presentations of what shapes are to be observed when waves are completely reflected from free or rigid boundaries (as well as dealing with partial reflection and transmission), students are rarely given adequate opportunity to sketch the shapes of reflected and transmitted waves under various circumstances. Without such opportunity, many students fail to develop significant understanding of the effects.²

Here again, asymmetric pulse shapes play a valuable role because they force the student to consider the sequence of events—a sequence that tends to be obscured in the symmetry of sinusoidal wave trains. Figures 9.5.1, 9.5.2, and 9.5.3 illustrate some of the points at issue. In the case of the transverse wave on a string, incident at a rigid wall (Fig. 9.5.1), the boundary condition of zero particle velocity at the wall requires a reflected pulse inverted in phase relative to the incident pulse. This is relatively clear to students, but what is far less clear is the fact that the reflected pulse must be a *mirror image*, inverted in phase, and this aspect becomes apparent only if an asymmetric pulse shape is employed as an example. The graphing exercises recommended in Section 9.3 pave the way for the thinking and visualization that are required.

²It is being taken for granted here that the discussion of reflections is preceded by adequate demonstration and discussion of the superposition of wave trains and pulses. Without prior consideration of superposition, the treatment of reflections is meaningless. Again, one of the best available treatments, with excellent illustrations, is that of the *PSSC Physics* textbook.

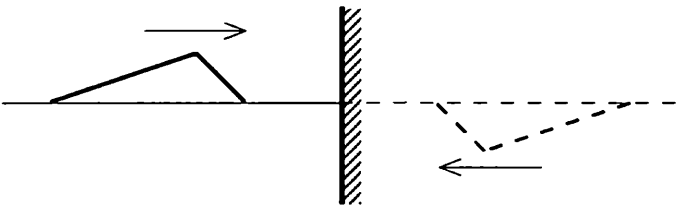


Figure 9.5.1 Transverse pulse on string is incident from the left at a rigid wall. To satisfy the boundary condition of zero particle velocity at the wall, the reflected pulse, imagined as originating in the fictitious region to the right of the wall, must be the phase inverted *mirror image* of the incident pulse.

Another important physical aspect of this situation is the fact that zero particle velocity on the string is maintained only at the rigid boundary itself. Elsewhere along the string the particle velocities are not zero as the reflected pulse propagates through the incident pulse. This is a point that eludes many students unless they are led to sketch the overlapping of the incident and reflected pulses at two or three successive stages and to consider the attendant particle velocities at various locations. The insights acquired here are enriched if students are led to return to the case of superposition of two symmetrical pulses passing through each other in opposite phase and in opposite directions, as illustrated in Fig. 9.5.2. At the instant the pulses “cancel” each other, the net deflection is zero all along the string, but the particle velocities are not everywhere zero. The particle velocity is zero in the central region where the maximum deflections cancel, but there are two regions of maximum particle velocity at the sides, one upward and one downward as shown.

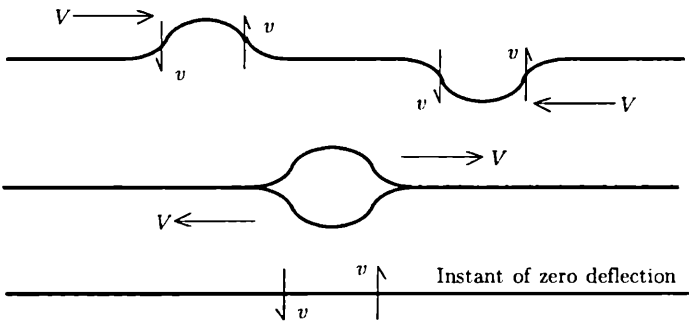


Figure 9.5.2 Superposition of two symmetrical pulses of identical shape but opposite phase, traveling in opposite directions at velocity V on a string. At the instant of coincidence of the peaks, the instantaneous deflection is zero everywhere, but the particle velocity v is zero only at the center and has maxima (one with upward and one with downward velocity) on either side of the center.

With this exposure, students are being prepared to understand the motions that will occur in standing waves resulting from reflection of sinusoidal wave trains.

In Fig. 9.5.3, we consider the reflection of a compression pulse (longitudinal wave) at a rigid wall. Here the boundary condition of zero particle velocity at the wall requires that the reflected pulse be a mirror image of the incident pulse without inversion in phase. If students have been exposed to the exercises recommended in Section 9.3, they are far better prepared to understand the reflection now under consideration.

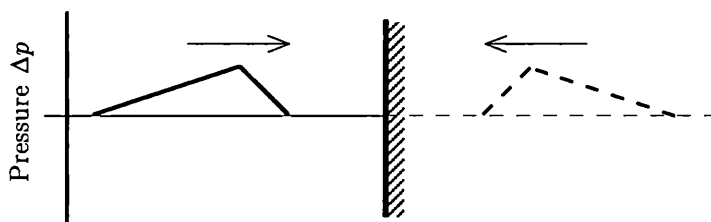


Figure 9.5.3 Compression pulse in a fluid (or on a slinky) is incident from the left at a rigid wall. To satisfy the boundary condition of zero particle velocity at the wall, the reflected pulse, imagined to originate in the fictitious region to the right of the wall, must be the *mirror image* of the incident pulse.

Homework and test questions should, of course, pursue variations on the examples given above: Incident negative transverse pulses and rarefaction pulses; incident wave coming from the right rather than from the left; free boundary instead of rigid boundary; sinusoidal wave trains instead of pulses; and also the reverse line of reasoning in which the shape of the reflected pulse is given and the shape of the incident pulse is called for. (As has been pointed out in earlier chapters, leading students to traverse a line of reasoning in both possible directions is highly conducive to learning and understanding. The reversal may seem trivial to an expert, but it is far from trivial to the learner.)

9.6 DERIVATION OF PROPAGATION VELOCITIES

The powerful, elegant, and rigorous way of showing that a disturbance will propagate under given circumstances with a velocity established by properties of the system is, of course, to apply the basic laws (mechanical or electromagnetic) governing the system, and to show that some form of wave equation will be obeyed. Such advanced treatment is clearly not appropriate or understandable in most introductory courses—except, perhaps, a few at the second year calculus-physics level. It is quite possible, however, to derive the wave velocity in a few interesting mechanical situations by applying only the requirements

of conservation of mass and the impulse-momentum theorem to a pulse that is *assumed* to propagate in a more or less steady state in one dimension. This is not nearly as rigorous an approach as deriving the wave equation, but it is quite reasonable and acceptable as a first cut at the problem.

An important advantage of such derivations is that they give students a chance to see basic laws, encountered earlier, actually employed in a powerful way to obtain significant results in new situations. So far, the only use students have seen for the basic laws (conservation of mass and Newton's laws of motion) has been in the highly restricted end-of-chapter examples arising in homework. Furthermore, students have not really seen general derivations using the basic laws; they have only seen examples of direct application to individual cases, used as separate exercises in homework.

Derivations of wave velocities are presented in the next three sections. I do not mean to recommend the introduction of these derivations in *all* introductory courses. They are best used at the discretion of the teacher. They might, for example, be made available to front-running students who would benefit from the deeper analytical insight, or they might be offered as opportunities for extra credit or independent study. In some college level courses, however, they are appropriate for an entire class, and, under such circumstances, they provide an opportunity for overview and synthesis that can come only from the spiralling back that is entailed.

The derivations outlined in the next three sections are certainly not new. They are presented in order to show how an essentially identical approach can be taken in three disparate cases in order to display the unity of the phenomena. It is this unity, and the continual cycling back to the same set of fundamental ideas, that make the intellectual experience impressive and lasting for the students.

9.7 VELOCITY OF PROPAGATION OF A KINK ON A STRING

Consider a string stretched horizontally under tension T as shown in Fig. 9.7.1. One end of the string is displaced abruptly, in the transverse direction, to a new position, and the corner or kink then propagates along the string to the far end. The action is assumed to be performed without change in the tension, and the angle θ between the original line of the string and the deflected portion is assumed to be small.

The basic (unproved) assumption is that the kink will propagate along the string at some velocity V determined by properties of the string. Given this assumption, we proceed to apply the restriction of conservation of mass and the impulse-momentum theorem to a small chunk of string (AB) just encompassed by the wave front in the small time interval Δt .

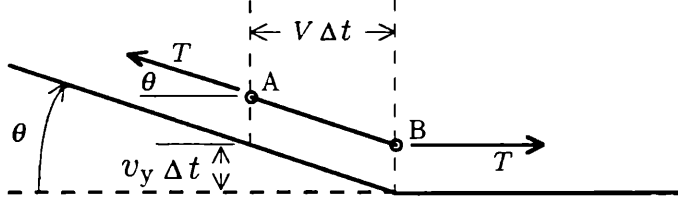


Figure 9.7.1 Taught string under tension T . End of string is displaced abruptly so that a kink propagates to the right along the string at velocity V . Transverse particle velocity is denoted by v_y . Force diagram is shown for segment AB.

The length of string encompassed will be $\Delta x = V\Delta t$, and, if we denote the mass per unit length of the string by μ , the mass Δm of segment of string AB encompassed by the wave in time interval Δt will be

$$\Delta m = \mu V \Delta t \quad (9.7.1)$$

This is, effectively, the continuity (or conservation of mass) equation and should be explicitly identified as such.

If we denote the particle velocity imparted to elements of the string by v_y , the change of momentum imparted to the chunk AB in the time interval Δt is $\mu v_y \Delta t$, and the impulse delivered in the y -direction is $T \sin \theta \Delta t$. Hence, by the impulse-momentum theorem

$$T \sin \theta \Delta t = \mu v_y V \Delta t$$

and

$$T \sin \theta = \mu v_y V \quad (9.7.2)$$

It is clear that one must now say something about the geometrical connection among θ , v_y , and V . From Fig. 9.7.1 it is apparent that

$$\frac{v_y}{V} = \tan \theta \quad (9.7.3)$$

[Those familiar with the formalism will note that, for a wave shape $y = f(x - Vt)$ propagating in the positive x -direction and obeying the small amplitude wave equation, the particle velocity v_y is given by

$$v_y = \frac{\partial y}{\partial t} = -V f'(x - Vt) = -V \frac{\partial y}{\partial x} \quad (9.7.3a)$$

which is the counterpart of Eq. 9.7.3]

Combining Eqs. 9.7.2 and 9.7.3 gives

$$V^2 = \frac{T}{\mu} \cos \theta \quad (9.7.4)$$

and, for small θ , with $\cos \theta$ close to unity

$$V = \sqrt{\frac{T}{\mu}} \quad (9.7.5)$$

the familiar expression for the propagation velocity of a small-amplitude wave on a string.

It should be noted that the small amplitude approximation not only takes θ to be small, but also ignores stretching and contracting of the string and any consequent small motions back and forth along the x -axis.

9.8 PROPAGATION VELOCITY OF A PULSE IN A FLUID

Figure 9.8.1 shows a tube of fluid in which a compression pulse is initiated by a rapid displacement of a diaphragm somewhere off to the left. It is assumed that the pulse travels to the right at a velocity V that depends on properties of the fluid. The initially undisturbed fluid has a pressure p_o , a density ρ_o , and a zero particle velocity ($u_o = 0$). The disturbed fluid behind the leading edge of the pulse has a pressure p , a density ρ , and a nonzero particle velocity u to the right. The cross-sectional area of the tube is denoted by A .

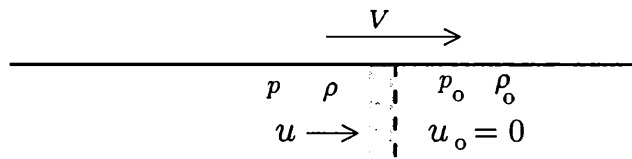


Figure 9.8.1 A compression pulse initiated off to the left propagates to the right in a tube of fluid. The cross-sectional area of the tube is denoted by A .

In a small time interval Δt , the leading edge of the pulse engulfs a mass of undisturbed fluid $\rho_o A V \Delta t$. In the region behind the leading edge, this material will be compressed into the volume $A(V - u)\Delta t$. Conservation of mass requires that

$$\rho A(V - u)\Delta t = \rho_o A V \Delta t$$

yielding

$$\rho(V - u) = \rho_o V \quad (9.8.1)$$

Since the net impulse delivered to the chunk of fluid engulfed in the time interval Δt is given by $(p - p_o)\Delta t$, and since the change of momentum of this

chunk is $\rho_o V u A \Delta t$, the impulse-momentum theorem requires that these two quantities be equal, and the result reduces to

$$p - p_o = \rho_o V u \quad (9.8.2)$$

Eliminating u from Eqs. 9.8.1 and 9.8.2 yields

$$V^2 = \frac{p - p_o}{\rho - \rho_o} \frac{\rho}{\rho_o}$$

or

$$V = \sqrt{\frac{\Delta p}{\Delta \rho} \frac{\rho}{\rho_o}} \quad (9.8.3)$$

and, from Eq. 9.8.2, the particle velocity u is given by

$$u = \frac{\Delta p}{\rho_o V} \quad (9.8.4)$$

It should be noted that none of the equations set down so far contain any small amplitude approximations. They are valid for large amplitude and are limited only by the steady-state assumption concerning the propagation. The change of state, although adiabatic, is inherently irreversible. The compression that takes place is not perfectly elastic; energy dissipation occurs in the process. The medium, after passage through the pulse and return to initial pressure level, has a higher thermal internal energy than it had previously. There is a net entropy increase associated with propagation of a finite amplitude pulse, and the final temperature is higher than the initial temperature on return to initial ambient pressure.

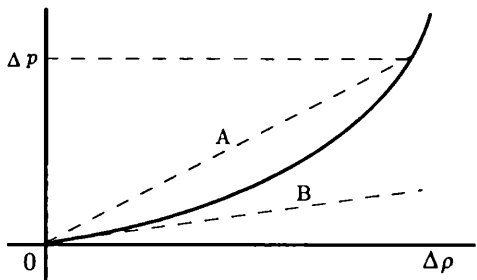
In the limit of small amplitude, Eq. 9.8.3 reduces to the familiar expression for acoustic velocity

$$V = \sqrt{\left(\frac{\partial p}{\partial \rho}\right)_S} \quad (9.8.5)$$

where the constancy of entropy S indicates reversible adiabatic compression in the limit of vanishingly small amplitude.

Equation 9.8.3 contains interesting physical implications and merits further examination. A graph of Δp versus $\Delta \rho$ for a fluid subject to large pressure changes would, in general, have the qualitative appearance shown in Fig. 9.8.2. The graph would be concave upward, indicating decrease in compressibility of the fluid with increasing pressure. Since, according to Eq. 9.8.3, the propagation velocity V is the square root of the slope of the chord (multiplied by a density ratio greater than unity) drawn in Fig. 9.8.2 from the origin to whatever Δp value is being considered, it is clear that higher compressions imply higher propagation velocities.

Figure 9.8.2 Schematic graph of Δp versus $\Delta \rho$ for any fluid subject to large adiabatic compressions and corresponding density changes. Line B is tangent to the curve at the limit of low amplitude; its slope determines the velocity of the acoustic pulse. The slope of chord A is $\Delta p/\Delta \rho$, which determines the propagation velocity of the finite amplitude Δp .



Thus, in a pulse initially having a rounded front such as that illustrated in Fig. 9.8.3, the higher pressure regions keep overtaking lower pressure regions ahead of them until the leading edge becomes the discontinuity in pressure and density called a “shock front.”³ As the higher pressure region overtakes the lower pressure ahead, it is also “runing away” from the lower pressure region behind. Thus, the spatial length (and the duration) of the pulse continually increase as the pulse advances.

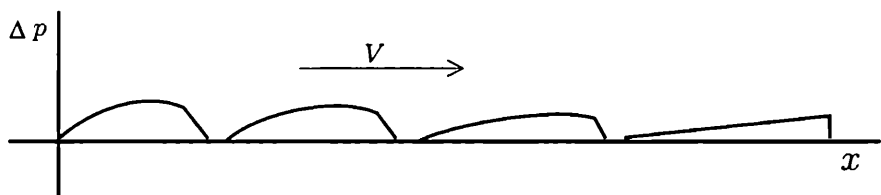
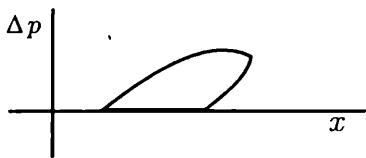


Figure 9.8.3 Formation of shock front and lengthening of pulse as an initially rounded wave pulse of finite amplitude propagates through the medium.

It is at the point of formation of the stable shock that the preceding equations become more rigorously applicable since, once the shock front is formed, the propagation is essentially steady except for the dissipation inherent in the finite amplitude transition. In other words, Eq. 9.8.3 is the fully correct equation for the velocity of a shock wave. Similarly, Eq. 9.8.4 is the fully correct equation for the particle velocity immediately behind a shock front.

Figure 9.8.4 Meaningless Δp versus x graph of a pressure pulse. Few students have the courage of conviction to recognize it as such.



As an illustration of the importance of leading students to confront what is *not* the case as well as what it is, consider the spurious pulse shape shown in Fig. 9.8.4. Very few students, even those in advanced courses, have the

³The shock is, of course, not a *mathematical* discontinuity. The region of pressure and density change has a width of the order of mean free paths of molecular motion.

courage to declare this diagram to be physically meaningless and impossible when it is first presented to them.

A note about a more advanced level for anyone who might be interested in pursuing it: With the introduction of an equation for energy conservation through the leading edge, one has a set of three equations that are called the “Rankine-Hugoniot Relations.” With the third equation, one can eliminate $\Delta\rho$ in terms of the independent variable Δp [analytically for an ideal gas and numerically for any substance whose equation of state data (p , v , T , and c_p) are tabulated] and thus obtain the propagation velocity directly in terms of the pressure amplitude. This, however, involves the use of thermodynamic relations well beyond the level of an introductory physics course.

9.9 PROPAGATION VELOCITY OF SURFACE WAVES IN SHALLOW WATER

The term “shallow” in the present context refers to a layer of water whose depth D is small relative to the wavelength of the surface gravity wave.⁴ Under these circumstances, the propagation velocity for small amplitude waves is given by the familiar, simple relation $V = \sqrt{gD}$. This relation will be derived below in a procedure exactly parallel to that applied in the two preceding sections.

The velocity relation indicates that the surface wave travels faster in deeper water, and I have, on various occasions, been asked by both students and colleagues how one can account for this physically. Given the greater inertia one naturally associates with deeper columns of water, it is quite reasonable to expect that the velocity would *decrease* with increasing water depth. The derivation gives insight into this nontrivial physical question.

Restricting the problem to one dimension as in the previous instances, consider a channel having a width Y and containing water to depth D (Fig. 9.9.1). Suppose that a wave pulse having a height h is generated by displacing a vertical wall in the channel somewhere off to the left of the figure. The basic assumptions are that the pulse propagates steadily with velocity V and that the vertical component of particle velocity behind the wave front can be ignored (i.e., that the particle velocity u can be treated as essentially horizontal). The water is, of course, treated as incompressible ($\rho = \text{constant}$).

Conservation of mass (continuity) equation: As the front advances for a small time interval Δt , the volume of initially undisturbed water passing into the disturbed region is $YDV\Delta t$ (sector ABCE in Fig. 9.9.1). After passing

⁴When the depth is large relative to the wavelength, the regime is entirely different and significantly more complicated. Under these circumstances, the pressure variations near the surface do not penetrate through the entire water column to the bottom of the layer. The mathematical analysis is much more sophisticated, and the velocity relation is a dispersive one. This regime will not be considered here.

through the front, this water is contained in the space $Y(D + h)(V - u)\Delta t$ (sector GHJC in Fig. 9.9.1) providing that the bottom of the channel is rigid and impermeable. Equating these two volumes gives

$$DV = (D + h)(V - u) \quad (9.9.1)$$

and Eq. 9.9.1 reduces to

$$hV = u(D + h) \quad (9.9.2)$$

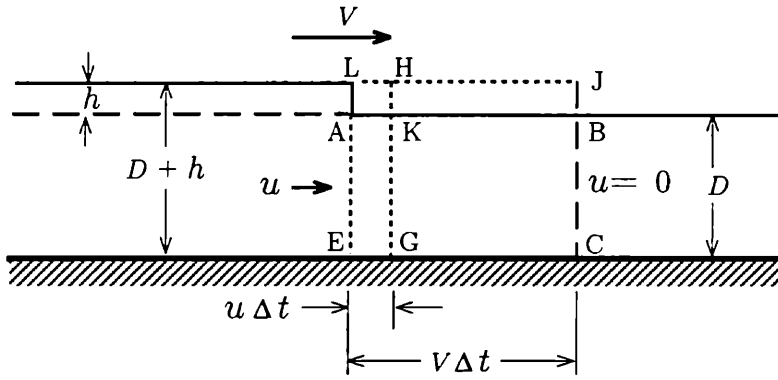


Figure 9.9.1 Positive surface wave pulse of height h propagates at velocity V along channel containing water of undisturbed depth D and density ρ . Channel width is denoted by Y and particle velocity behind the wave front by u .

Impulse-momentum equation: The initial hydrostatic pressure distribution through the depth of the fluid plays no role in acceleration of fluid particles since the attendant forces are balanced throughout. The particle velocity u is imparted by the *unbalanced* force applied over the area YD of initially undisturbed fluid by virtue of the excess pressure ρgh , which penetrates the entire depth of the water column. (This is where the restriction to long waves on shallow water enters this derivation.) Thus, an impulse $\rho ghYD\Delta t$ is imparted to the mass of water that passes through the front. The momentum change of this water is $\rho YDVu\Delta t$. Equating the impulse and the corresponding momentum change yields

$$gh = Vu \quad (9.9.3)$$

and, eliminating u from Eqs. 9.9.2 and 9.9.3, gives

$$V^2 = g(D + h) \quad (9.9.4)$$

which, in the limit of small amplitude, gives the familiar

$$V = \sqrt{gD} \quad (9.9.5)$$

One can now begin to see how it is that the wave travels faster in deeper water. First let us rearrange Eq. 9.9.2 as follows:

$$uD = h(V - u) \quad (9.9.6)$$

The two volumes in Eq. 9.9.6 are interpreted in Fig. 9.9.1: The left-hand side refers to the volume of water in the sector AKGE (without the Y and Δt terms) while the right-hand side refers to the volume in the sector HJBK (also without the Y and Δt terms).

Equation 9.9.6 says that, in the time interval Δt , a volume of water $uDY\Delta t$, initially ahead of the wave front in AKGE, must be transferred into the region $h(V - u)Y\Delta t$, now behind the wave front in HJBK, as shown in the diagram. (This is not to say that the first volume is literally lifted into the location of the second. The actual trajectories of the water particles are not being mapped out in this highly restricted analysis.) It is evident from Fig. 9.9.1, however, that the continuity equation requires that the wave front advance far enough in any given interval of time Δt to accommodate the volume $uDY\Delta t$. This means that the front must advance *farther* if the water is *deeper*.

This analysis yields the unusual insight that, in the case of the surface wave in shallow water, the progress of the wave is, in an important sense, driven by the continuity requirement rather than by a force. A force, resulting from the pressure difference provided by the wave height, is, of course, necessary in order to generate any disturbance at all, but the velocity is then dictated by the continuity requirement since, in a given time interval, the front must advance farther in deeper water in order to conserve the shifted mass.

Returning to Eq. 9.9.4, we note that, if we admit appreciable wave height h relative to depth D , higher amplitude regions of a pulse propagate at higher velocity than lower amplitude regions and that the pulse therefore tends to build up into a shock front, as does the finite amplitude compression pulse in a fluid. In the case of the water wave, however, the wave “breaks,” the most familiar illustration being the running up of waves on a beach. Because of the analogy between finite amplitude surface waves on water and finite amplitude pressure waves in a fluid, some experimentalists have used water waves to model and study the nonlinear effects associated with the intersection and reflection of shock waves.

Final comments: The three preceding derivations of wave propagation velocity have been deliberately designed to dramatize the power of an identical approach applied to seemingly very different physical situations. The results reveal the profound unity underlying the disparate phenomena. This is invariably deeply impressive to students whose previous experience has not exposed them to a comparable synthesis. The repetitive application of the basic laws also plays a very important role in helping the students master the reasoning rather than just memorizing a procedure in an isolated instance.

Another virtue of these derivations is that they entail a very strong physical content, and requirement for phenomenological reasoning, with a minimum of mathematical complexity. Better students, especially those heading for physics or engineering, are very much in need of exposure to such reasoning early on. Many of them come to advanced courses (or even to graduate school) without having had the experience, and this becomes very evident in their performance. Faculty complain bitterly about the inadequacy of the students without identifying the source.

9.10 TRANSIENT WAVE EFFECTS

Once waves and wave pulses have been studied and thought about for a while, student perspectives can be significantly broadened by leading them to reconsider, on a purely qualitative level, a variety of everyday experiences as well as some of the physical situations dealt with in problems earlier in the course. Many of these situations involve transient wave phenomena, but very few students discern this spontaneously since most of the phenomena must be visualized without being directly seen. Students do begin to comprehend the connections, however, if led into thinking about these effects and visualizing them. This, in turn, propels them into a more sophisticated approach to phenomena that transcend direct sense experience. Some examples:

- 1 If we lay a long rod down on the table and displace a box at the far end of the rod by pushing on the near end, what is it that happens at the very beginning of action? Is the force exerted on the rod at the near end equal to force exerted by the rod on the box? Does the displacement of the box begin at the same instant that we push on our end of the rod? What is it that happens in the rod and at the box before a steady state is attained?
- 2 An Atwood machine (or any other system of objects and strings dealt with in earlier dynamics problems) is initially held stationary and is then let go and allowed to accelerate. What happens between the instant of letting go and the achievement of the state of steady acceleration? In light of these examples, what do you suppose happens in elevator cables?
- 3 Consider the bouncing of a ball from a floor or a wall. What happens in each of the interacting objects during the collision? What happens when two balls collide? What is the origin of the sound we hear when collisions occur?
- 4 What happens when a balloon bursts? What is the origin of thunder?
- 5 What happens in a crank shaft when a torque is suddenly applied at one end?

- 6 You push horizontally at the top of a large box and slide it along the floor against the opposing frictional force. What happens between the application of the push and the time at which the box begins to slide?

It is not usually necessary to supply a long list. Once students are cued to such questions, many of them can begin to invent questions of their own, and many of these turn out to be interesting and ingenious. The object is to get them to discern the presence of longitudinal, shear, and torsion waves in virtually every dynamic interaction that takes place around us—despite the fact that these waves elude direct sense perception.

9.11 SKETCHING WAVE FRONTS AND RAYS IN TWO DIMENSIONS

Although most textbooks exhibit diagrams showing transmission, reflection, and refraction of waves in two dimensions, and although excellent film loops are available showing these effects in ripple tanks, the students encounter few, if any, homework problems that lead them to sketch ray and wave-front diagrams of such phenomena themselves. The consequence is that many students are unable to sketch correct diagrams of their own in either representation, especially if the orientation of the interface is changed from that of an illustration given in the text.

This is not a formidable conceptual problem, and most students grasp the ideas fairly quickly. What is missed in instruction is the fact that practice is essential regardless of the clarity of text diagrams and discussion. Numbers are not necessary; purely qualitative questions form the best vehicle.

Questions should include aspects such as the following: Given an interface at which propagation velocity changes and given an incident ray, sketch the reflected and transmitted rays, and then sketch the same situation with corresponding wave fronts instead of rays. Changes in wavelength should be indicated in the wave-front diagram. Students should then be led to ring their own changes on this theme: Orient the interface differently on the page; reverse the velocity change (i.e., if the initial question had the incident wave in the faster medium, put the incident wave in the slower medium); start with the wave-front diagram and then sketch the corresponding ray diagram; start with the transmitted (or the reflected) rays or wave fronts as given and sketch the rest of the diagram. (Note that students should be led to traverse the line of reasoning in all possible directions, backwards as well as forwards.)

Such situations are usually first encountered with ripples. The sketching should be repeated in the encounter with light. As has been pointed out repeatedly, it is the opportunity to use an idea, after elapse of time and in an altered context, that leads to mastery and retention. In many students, retention is feeble on only one exposure.

The sketching of such diagrams should be required on tests as well as in homework. The phenomenology is important, and since these ideas are readily mastered with a little practice, those students who have done their homework can do well on the test questions, thereby acquiring reinforcement that is frequently lacking in much of our testing.

9.12 PERIODIC AND SINUSOIDAL WAVE TRAINS

Although most texts give reasonably adequate developments of the basic relation

$$V = \nu\lambda \quad (9.12.1)$$

connecting propagation velocity, frequency, and wavelength for periodic wave trains, many students proceed to use it blindly, as a memorized formula, with no understanding of its origin or justification. Being able to substitute in the formula and obtain correct numerical answers is no indication of understanding.

The learning problem here is a very basic one: There is a profound reluctance among many learners to go back, on their own initiative, to the operational definitions underlying a line of reasoning, articulate the definitions, and use them to obtain a result or draw an inference. They will gloss over text presentations and avoid the sequence of definition and reasoning unless explicitly required to pursue it. (See Sect. 6.13 for another example.)

Since, in this instance, the definitions are simple and the reasoning purely arithmetical, this is a valuable opportunity for practice that helps lower resistance to such reasoning. Students should be led to respond to questions such as: (1) State in your own words the definitions of the quantities represented by the symbols V , ν , and λ . (2) Explain in your own words, with accompanying diagrams, how these definitions lead directly, through purely *arithmetical* reasoning, to relations such as $V = \nu\lambda$, or $\nu = V/\lambda$, or $\lambda = V/\nu$.

The representation of sinusoidal wave trains presents not only an opportunity to exploit the various forms of Eq. 9.12.1 but also an opportunity to cycle back to radian measure and the reasons for invoking it (see Sects. 1.13 and 4.6). The great majority of students, even among those in engineering-physics courses, have great difficulty grasping the need for the $2\pi/\lambda$ factor in

$$y = A \sin \left[\frac{2\pi}{\lambda} (x \pm Vt) \right] \quad (9.12.2)$$

They should be led to explain the necessity of this term in their own words.

Finally, an effective kind of homework (and test) question missing from most textbook collections is exemplified in the following (note that only qualitative sketching, not numerical plotting, is being called for):

Given the “photograph” (i.e., form at a given instant of time) of a sinusoidal wave train, sketch the photograph that would be obtained if

- (a) the amplitude and the frequency were doubled while velocity remains unchanged.
- (b) the frequency and velocity were both doubled while the amplitude remains unchanged.
- (c) the wavelength and amplitude are reduced by a factor of three while the velocity is doubled.
- (d) and so on and so forth.

Students might then be invited to make up, for themselves, a parallel series of questions in which one starts with the y versus t graph at a fixed point in space instead of with a given drawing. Students are rarely required to make up questions and problems of their own, and, because of lack of practice, the majority are extremely reluctant to do so. Simple cases of this variety therefore constitute a valuable opportunity, affording practice that helps lower resistance.

Still another aspect with which many students have difficulty is the perception that, in transmission and reflection of sinusoidal waves at an interface, it is the frequency and not the wavelength of the wave that is preserved. They must be led to visualize the arrival of crests and troughs at the boundary and to visualize what effect this has on the disturbance transmitted through the boundary. Such exercises are easily coupled to the ones described above and to those suggested in Section 9.11.

9.13 TWO-SOURCE INTERFERENCE PATTERNS

Understanding of two-source interference patterns does not emerge from passively viewed demonstrations or from numerical exercises with the formulas connecting wavelength, spacing between the sources, and the angle between loci of constructive or destructive interference and the principal axis. This is not to say that the demonstrations and the numerical problems are superfluous; both are essential to learning and understanding. The point is that other views are also necessary. It is the overall mixture that induces understanding.

The additional ingredients that are helpful, but that are rarely included in exercises provided by the textbooks, involve verbal and qualitative insights. It is effective, for example, to have students look at an actual ripple tank pattern (or a clear picture of one) and confront the question: What is happening at this particular point in the pattern? They must decide, by inspection of the pattern, whether the arbitrarily chosen point is at a node, an antinode, or somewhere in between, and also discern the order of the location. Although

this is an exceedingly simple question, it is not trivial, and many students initially have difficulty because they have never been led to examine the pattern from this point of view. The reversal of this question is equally important: Point to a location that lies between a second order constructive and a second order destructive interference. Such questions make useful, effective, and easily graded test questions (now that it is easy to reproduce pictures of this kind).

A next level resides in having students make explicit connection between the ripple pattern (usually the first encountered) and analogous acoustic and optical patterns. Textbooks tend to assume that the connections are obvious, but this is not the case for many students, and it is important to lead them into articulating the similarities and differences. For example, as a student looks at the ripple pattern, one can ask what he or she would have to do to discern a corresponding acoustic pattern: Where would you go? What would you expect to hear at locations that correspond to such and such points in the ripple pattern? How would you map out loci of constructive and destructive interference?

The parallel questions should be raised with optical patterns. Many students fail to connect the ripple pattern (always viewed as a whole from above) with the bright and dark regions of a two-slit optical pattern projected on a screen. They should be led to visualize where they would have to stand and what they would see if the ripple pattern were an optical one in which nothing could be seen from above.

The optical pattern should also be explicitly connected to the ripple pattern. Now that we so glibly show two-slit optical patterns by using lasers, it is easy to exhibit the entire two-dimensional constructive and destructive interference pattern, making it visible from the side by sprinkling chalk dust into the region between the slits and the screen. [Some teachers apparently do this chalk dust demonstration with a grating and imply that the pattern corresponds to the two-source pattern in the ripple tank. It happens that the grating and two-slit patterns are virtually undistinguishable qualitatively in the laser demonstrations. Unfortunately, this is very misleading. The grating effect differs from two-slit interference in a profound and significant way, and a grating pattern in the ripple tank does not look anything like the two-source pattern (see discussion in the following section).]

9.14 TWO-SOURCE VERSUS GRATING INTERFERENCE PATTERNS

Mathematical analysis, showing the enormous sharpening of the regions of constructive interference, provides the compelling justification for using gratings instead of double slits in spectrometers. Many students, however, fail to grasp the full force of the mathematical analysis if it is invoked, and, in many

courses, the mathematical analysis is inappropriate because the students are not ready for it. It is therefore important to try to develop the difference between these two situations phenomenologically. Unfortunately, modern laser demonstrations tend to conceal the physical difference, and it is advisable to resort to older approaches.

In older textbooks, students were usually given exercises that involved drawing the Huygens wavelets emerging from each opening when a plane wave is incident on a grating and then identifying the plane wave fronts that are reconstituted in the emerging beam. Such exercises were valuable except that the books rarely made sufficiently explicit the contrast between the pattern formed by the overlapping circles in the two-source case and the plane wave fronts going off in very sharply defined directions after transmission (or reflection) with a grating. Few modern texts lead students to perform and interpret such drawings, and the physics of the grating effect gets lost through oversight. This is unfortunate because the concept involved is subtle and important.

For this reason, it is well worth leading the students to draw the patterns with compass and straight edge and to describe in their own words the essential differences between the two-source and the grating patterns. It is also well worth setting up a grating pattern in the ripple tank. Although this is a bit tricky, it can be done. One needs to use a low frequency (wavelength of about 2 cm) and a barrier grating with six or eight openings of similar width.

It is necessary to be prepared for what to look for in the ripple tank, since the grating pattern is completely different from the two-source pattern. When the grating pattern is stabilized, one sees only a network pattern, a crisscrossing of straight waves—nothing like the “fingers of a fan” associated with two sources. If one lets the eye sweep with the straight waves in the grating pattern, one begins to see three sets of straight waves (it is difficult to get anything beyond first order): One set propagates in the original direction along the principal axis, and two sets go off on either side of the principal axis in the direction of first order constructive interference. The fact that the constructive interference gives straight waves (not curved ones) moving in a sharply defined direction is the qualitative indication of the sharpening produced by the grating. One receives no signal from any direction slightly different from that in which the straight wave is propagating. With ordinary light, the plane waves emerging from the grating must, of course, be focussed by a lens in order to form an image on a screen. (This focussing is performed for us by the lens of the eye when we look through a grating.)

These are some of the physical insights that get lost in many modern demonstrations relying on the laser. This is said not with the intent of deprecating the use of lasers in optics demonstrations: The device has made important demonstrations more visible and easier to perform by many orders of magnitude. It is tremendously valuable. One need only be careful about those conceptual issues that might be glossed over or confused.

9.15 YOUNG'S ELUCIDATION OF THE DARK CENTER IN NEWTON'S RINGS

An illustration of how an element of historical knowledge can enhance physics teaching resides in examining Thomas Young's modification of the well known Newton's rings experiment—a modification that lent powerful support to the wave model of light at a time of vigorous controversy over the wave and particle models. The modification, at the same time, exhibits to the students a fine illustration of a crucial experiment.

One troublesome feature of the attempt to explain Newton's rings with a wave model was the fact that the reflection pattern for a film of air between glass boundaries was observed to be dark rather than bright at the center—where the film thickness is essentially zero and where the incident and reflected waves would interfere constructively rather than destructively if the phase difference between them were zero.

It is easy now, in retrospect, to assert phase inversion at one interface, but, at the time, this was a substantial puzzle. The entire problem is a subtle one for students new to the study of physics. First, it takes time and thought for them to begin to appreciate the fact that there is indeed a nontrivial problem and that something needs to be explained. Furthermore, they are still very vague as to what is “waving” in the case of light, and, since the connection must be made to mechanical situations by analogy rather than by direct sense experience, the reasoning is subtle and requires thought and discussion for genuine understanding.

Young, who also had no notion of what was “waving,” pursued the analogy to mechanical behavior and surmised that the undulations, whatever they might be, kept the same phase or were inverted in phase depending on whether they were reflected from “less dense” or “more dense” interfaces, something like waves at free and fixed boundaries of strings.

To test this hypothesis, he utilized a lens of crown glass having an index of refraction of about 1.5 and a plate of flint glass having an index of about 1.7. Between the glasses he placed, instead of air, a film of sassafras oil, selected because it had an index of refraction falling between those of the two glasses. Young predicted that, if the wave analogy was correct, the reflected ring pattern should now have a bright, rather than dark, center. He reasoned that both reflected waves would either retain the same phase or both be inverted in phase (relative to the incident wave) since both reflections took place at interfaces at which the reflecting medium was optically more dense than the incident medium. Young showed that the center of the ring pattern, under these circumstances, is indeed bright, and this lent powerful support to the wave model.

This story provides the students an opportunity to engage in qualitative physical thinking of great importance and in a very rich context, one linking their still shaky grasp of mechanical wave reflections with growing insight into

the nature of light. It also provides an excellent illustration of the design of an experiment to test a hypothesis as well as an opportunity to sharpen the ability to distinguish between observation and inference. Furthermore, although some writers have denied the existence of really “crucial” experiments, it seems to me that this investigation comes very close to satisfying the classical criteria for a crucial experiment.

9.16 SPECULAR VERSUS DIFFUSE REFLECTION

Although some textbooks present clear diagrams and emphasize the distinction between specular and diffuse reflection, many gloss over these concepts and their attendant physical effects much too quickly and casually. Apparently the ideas are deemed too simple to be worth valuable time and space. Yet many students, even when using books that do not shortcut the concepts, emerge without having absorbed their significance. This is probably due to the fact that, even when the effects are described, qualitative questions about them rarely arise in homework or on tests. Assigned problems tend to concentrate on numerical or geometrical determination of image positions with mirrors and lenses, and students are rarely impelled to visualize the overall array of physical phenomena, especially those having to do with light radiated or reflected from the object.

Probably the most significant gap that develops in this connection is the failure of students to become explicitly conscious of the fact that all nonself-luminous objects that we see are seen by diffuse reflection of ambient light and that each point on the object acts as a point source, reflecting light in *all* directions. As a result, very few students are aware of any connection at all between mirrors on the one hand and ordinary objects (nonmirrors) on the other.

Like many other phenomena impinging on immediate experience (e.g., friction; thermodynamic equilibrium in the surrounding air), diffuse reflection is very complex. It is not isotropic and varies significantly with the nature and angle of incident light. One need not spend excessive amounts of time on such matters, but helping students become aware of the fact that most of what we see involves reflection in the domain between purely diffusive and specular, seems a worthwhile extension of their view of the nature of “seeing.”

Similarly, many students fail to recognize that each point on a luminous source is also radiating light in *all* directions. Many diagrams that students see in texts (especially science texts at pre-high school level) are incorrect or, at best, misleading since they are likely to show only special rays in preferred directions without showing that a special selection is made out of an infinite bundle. Since students are rarely, if ever, asked to sketch diagrams in which they themselves show the multiplicity of rays emerging from each point on the object and then select a preferred ray for further tracing, they end up without firm assimilation of the concept.

The consequence of this widely implanted gap is that few students start a ray diagram involving image formation by mirrors or lenses with explicit recognition that each point on the object is an independent point source of a spherical wave front and an infinite bundle of rays. Eventually, many misconceptions regarding images and image formation can be traced back to this gap as at least one of the root causes.

In the initial stages of dealing with ray diagrams, it is advisable to lead students to sketch the spherically divergent bundle of rays from an object point and only then to select the special ray that will be readily traced through the system. Only after the concept has been firmly registered should the clutter of unneeded rays be dispensed with.

Students should then be led to see the difference between a mirror and an ordinary object, that is, the difference between specular and diffuse reflection.

9.17 IMAGES AND IMAGE FORMATION: PLANE MIRRORS

Goldberg and McDermott (1986) have studied student misconceptions regarding images and image formation by plane mirrors. They show that certain misconceptions are held by large numbers of students and that these misconceptions persist through exposure to conventional instruction, which concentrates on quantitative aspects and fails in leading students to deal with, and interpret, the qualitative observations and phenomena.

Goldberg and McDermott report student performance on four tasks. In Task 1, students were asked to put a finger at the location of the image of a vertical rod they saw in a plane mirror. The majority of preinstruction students (65 to 75%) responded correctly by locating the image behind the mirror while 20 to 30% located it on the mirror. (The latter response is found to be common among individuals encountering the question for the first time.) The great majority of postinstruction students (95%) located the image behind the mirror. It is apparent that this aspect of instruction does register fairly firmly.

In Task 2 the student was asked to keep the finger on the image position while considering the following question: "Suppose you were sitting where I (the interviewer) am now, about two feet to your left, and I asked you to put your finger above the image. Would you put your finger at exactly the same place it is now or at a different place?" About 45% of the preinstruction students responded correctly, namely that the image would remain in the same place, while the majority (about 55%) responded that the image position would change. Most of the latter responses were based on the supposition that the image would lie along the line of sight between the viewer and the object (in these interviews, the object was placed closer to the mirror than either of the two viewers.) Among the postinstruction students, 70% answered

correctly while 30% still maintained that the position would change. Many of the latter changed their view, however, as the interviewer asked them to draw their own ray diagrams; those who drew correct ray diagrams were able to revise their prediction.

In Task 3 the mirror is kept covered during the interview. The student is seated in a position well beyond the right edge of the mirror, and the vertical rod is also placed beyond the right edge in a position such that the line of sight from the student to the rod intersects the mirror. (No light from the object, however, can be reflected to the student's eye.) The question is then asked "If we uncover the mirror, would you see an image of the rod?" After this question is answered, the interviewer asks "Would I see an image of the rod?" (The interviewer is seated to the left of the student in a position where light from the object would be reflected to his eye.) Among the postinstruction students, 70% gave the correct response (no/yes) while 5% said "yes/no" and 25% said "yes/yes." Virtually all the students who gave the last response used one kind of reasoning to predict what they themselves would see and another to predict what the interviewer would see. They erroneously decided that they would be able to see the image because it would be on the line of sight to the rod. On the other hand, by correctly applying the law of reflection, they concluded that the interviewer would be able to see the image.

These investigations show that incorrect views and interpretations persist beyond conventional class instruction with large numbers of students—even in the relatively simple situation involving plane mirrors. Goldberg and McDermott also conclude that instruction that explicitly raises phenomenological questions such as those used in the interview tasks is markedly effective in improvement and retention of understanding.

My own experience coincides with that reported by Goldberg and McDermott. I have also observed that one of the difficulties behind the off-the-cuff incorrect responses of the students is the gap discussed in Section 9.16: Few students explicitly invoke, in thinking about the phenomena involved, the idea that each point on the object sends out rays of light in all directions and that the image in the plane mirror does exactly the same thing for all those rays (from the object) that strike the mirror. Thus they do not, on exposure to conventional instruction about plane mirrors, begin to form a clear conception of what is meant by "virtual image." This concept offers still greater difficulty to many students when it arises in the more subtle contexts of lenses and curved mirrors. Stronger emphasis on the virtual image concept in the case of plane mirrors greatly facilitates later learning. Just invoking the name, however, is not enough; students must be helped to articulate the ideas in their own words after looking into real mirrors and drawing their own diagrams, and they should be led to sketch *entire bundles* of rays, not just isolated special rays.

9.18 IMAGES AND IMAGE FORMATION: THIN CONVERGING LENSES

Goldberg and McDermott (1987) report an investigation, similar to that described in Section 9.17, of student conceptions regarding image formation by converging lenses. All the tasks involved questions concerning events on an optical bench that the interviewer and student viewed from the side. The apparatus on the bench included (1) an object consisting of the luminous horseshoe-shaped filament of an unfrosted light bulb; (2) a converging lens (diameter 7.5 cm; focal length 17 cm); and (3) a translucent screen. The inverted image of the filament was focussed on the screen, and the student viewed the screen from the side facing the lens.

In Task 1 Goldberg and McDermott framed the following question: "If the lens is removed, leaving the object and the screen where they are, would anything change?" The majority of preinstruction students (60%) responded that the image on the screen would become erect. About 40 to 45% of postinstruction students (who had received conventional instruction in geometrical optics) gave the same response. On further investigation, Goldberg and McDermott found that the response in this task was, to some extent, confounded by the sharp nature of the object (the luminous filament); fewer students gave the incorrect response when an obviously diffuse source (a frosted light bulb) was used as the source. The overall results, however, indicate that large numbers of students, even after instruction, fail to recognize the absolute necessity of the lens for image formation. Even though they are fully aware that images of objects in the room do not form on the walls, they do not invoke such everyday experience when viewing the apparatus on the optical bench.

In Task 2 the interviewer holds a piece of opaque cardboard above the lens but does not cover any part of it. The following question is then asked: "Suppose I were to bring this cardboard down and cover the upper half of the lens, leaving the lower half uncovered, would anything change on the screen?" Between 90 and 95% of pre-instruction students predicted that half of the image would vanish, and between 55 and 75% of post-instruction students made the same prediction. Only a minority of the latter recognized that the entire image would remain and become less bright. In other words, even after instruction, only a minority of the students recognized that any portion of the lens is capable of forming the image.

The misapprehension arising in this context stems largely from the routine way in which students learn to draw ray diagrams. The texts show only the principal rays in order not to clutter up the diagrams. The students never draw anything but the principal rays; they never show the divergent bundle of rays emitted from every object point (cf. Sect. 9.16). They are never led to visualize what happens to the bundle that passes through any part of the lens regardless of where the principal rays happen to be. When they look at a diagram showing only the principal rays, they see the cardboard as cutting

off the principal rays in the upper half of the diagram and conclude that half the image must disappear.

If one presents students with a problem sketched in such a way that the arrow representing the object is taller than the converging lens, many students find it impossible to draw the principal ray parallel to the principal axis because it never intersects the lens shown in the diagram. They fail to recognize that, once the plane of the lens is given, all the principal rays can be drawn regardless of whether or not all the principal rays actually pass through the lens. This reflects a gap in understanding closely related to that in the interview task.

In Task 3, starting with the image focussed on the screen, Goldberg and McDermott invoked the question: "Suppose I were to move the screen toward the lens. Would anything change on the screen?" Among the post-instruction students only 35 to 40% recognized that the image would become fuzzy and disappear; the remainder expected the image to persist, changing in size and perhaps becoming somewhat fuzzy. Goldberg and McDermott report the following:

Many of the remarks made by the students indicated that the function of the screen was widely misunderstood. Often they did not think of it as a diffuse reflector or transmitter that, when located at a particular position for a given object distance, makes it possible for an observer not looking along the axis of the lens to see the image. Instead they seemed to believe that an image can be seen on a screen no matter where it is placed along the axis. In some cases this claim seemed to be buttressed by a misinterpretation of the experience of watching someone else use a slide projector. The students may have remembered that, in order to make the image larger, the screen had to be moved further from the projector. However, they did not recall that it was necessary simultaneously to refocus the projector by changing the object distance.

Many students fail to recognize the significance of the ray diagram in the sense that the intersection of the principal rays determine a unique image position and that an image is therefore not formed in any location other than the image plane. Furthermore, students have never been led to interpret the role of the screen in terms of diffuse reflection, that is, they have not been led to show each image point on the screen as a source of a divergent bundle of rays (again cf. Sect. 9.16).

In Task 4, the student, while looking at the image on the screen, is initially asked if he or she would still be able to see the image from his or her present position (on the side of the screen facing the lens) if the screen were removed. Virtually all the students recognized that the screen was necessary for reflecting the image so that it could be seen, although several of the preinstruction students volunteered that they might be able to see the image on the wall

several meters beyond the screen.

The students were then asked whether they would be able to see the image if the screen were removed and they were free to take up any other position they wished. Goldberg and McDermott report:

Many of the students especially the prestudents, seemed to have difficulty in understanding the question. Their remarks indicated that they could not conceive of an image as existing in free space, independent of a surface. The majority of the students, both pre and post, said either that they would not be able to see an image or that they might be able to see an image if they could place their eye at the screen position. Several of the students who gave the latter response seemed to think of their eye as simply replacing the screen.

At this point in the interview, the investigator actually removed the screen and directed the student to move about two meters beyond the initial screen position and to look along the lens axis toward the lens. With this guidance, almost all of the students were able to see the aerial image. Many appeared surprised that they were able to see anything.

When the investigator asked for the location of the image, only a very few students were able to state correctly that the image was located at the same position that the screen had been. The rest of the students gave a variety of answers. The prestudents, especially, tended to say that they thought the image was at or in the lens.

It is my own experience that, if one asks students to give an overall description in their own words of what happens to the light which emanates from the object, is intercepted by the lens, and ends up forming the image on the screen, a substantial number tell a story that effectively boils down to something like "The image travels from the object to the lens; in the lens it is turned upside down; then it travels to the screen."

Again in my own experience, I have found it helpful to require students, when drawing ray diagrams on homework and tests, to write out full verbal descriptions of how each principal ray is drawn (this in connection with both converging and diverging lenses.) Many of the initial versions tend to be gibberish, indicating that the students have not registered the ideas involved and are just following memorized diagrams or procedures without making explicit connection to the definitions of principal foci and rays. Only after the verbal description has been written out correctly and clearly at least once (preferably in connection with a problem that involves image formation by two lenses in sequence) can this requirement be relaxed.

Another very useful technique for tests and homework is to help the students reverse the direction of reasoning. Instead of asking for the image loca-

tion given an object position, one can give the image and lens locations and ask where the object must have been. (I know of very few texts that ask for diagrams under these circumstances.) Still another version, of course, is to ask for the lens position given object and image.

Generally speaking, very little genuine understanding of images and image formation is registered under homework that consists principally of numerical calculations utilizing either the Gaussian or Newtonian lens equations. This remark is not meant to advocate elimination of the numerical work. The latter is necessary and important, but it needs to be coupled directly, preferably in the same problems, to the phenomenological aspects that are shown to be far more difficult to assimilate than the numerical routines.

It is clear that homework assignments, tests, and *laboratory work* should contain questions and problems of the type used in the tasks described by Goldberg and McDermott. Many variations on these specific tasks could and should be devised. One cannot avoid the conclusion that most conventional instructional routines fail to lead many students to the kind of understanding we would like them to achieve.

9.19 NOVICE CONCEPTIONS OF THE NATURE OF LIGHT

Just as students come to the study of motion with many deeply rooted, commonsense preconceptions, they also come to the study of light with a variety of preconceptions. Many of these preconceptions go unnoticed by teachers and texts and interfere with the development of understanding of the physics. Watts (1985) summarizes some of the notions held by novices. (Not every individual of course, simultaneously holds all the views described; percentages vary for different groups and different age levels.)

Relatively few students hold a conception of light as a physical effect existing apart from its sources and effects. Light illuminates objects so that they are seen, but the act of seeing is not explicitly associated with the arrival of light at the eye of the observer. (I have observed that a few students even reinvent the ancient idea proposed by Parmenides that something emanates to an object from our own eyes to make seeing possible.) Sources of light (such as lamps or candles) are “seen” at a distance but are not thought of as sending out light.

Reflection is something that happens with mirrors but not with a sheet of paper, walls of a room, or other objects. (See Section 9.16.) Color filters are seen as *adding* color to white light.

It is important for a teacher to be aware of the fairly wide incidence of such preconceptions and to help students unsettle them through appropriate questions leading to encounter of contradiction and inconsistency. Rapid assertions of the “correct” view do little good.

9.20 PHENOMENOLOGICAL QUESTIONS AND PROBLEMS

Preceding sections have indicated the importance of giving students qualitative, phenomenological questions about both mechanical waves and optical phenomena in order to strengthen their intuition and build firmer understanding of both the phenomena and the underlying concepts that we invent. One very powerful type of question is the “What will happen if . . . ?” variety that cultivates hypothetico-deductive reasoning. Most of the researches that were cited above used such questions as revealing probes, and the very questions that have been quoted cry out to be incorporated in day to day instruction.

In addition to the illustrations that have been provided in geometrical optics, it is desirable to add similar questions in various aspects of physical optics. For example, in addition to making calculations of wavelength from an interference pattern (say with parallel fringes formed by a wedge-shaped film between glass plates), one might ask students to sketch how the pattern would change if one of the plates were to be displaced parallel to itself, either thickening or thinning the film; sketch how the pattern would change from the one observed if the two plates did not have the same index of refraction; sketch how the pattern would change if the color of the incident light were changed from red to green. Such changes are rarely demonstrated in connection with discussion of patterns formed by thin films, and even if demonstrated, the ideas do not register unless students are asked to sketch the effects for themselves. (This also applies to changes in patterns formed by slits and gratings. Although these are usually demonstrated, many students fail to produce correct sketches of their own unless led to practice. The tendency is to try to memorize what was shown rather than to reason it through to the underlying principles.)

The brightness of an image formed by a lens invites return to ratio reasoning and scaling without substitution in a formula (see discussion in Chapter 1). The combined effects of focal length and aperture of the lens offer many students, even those in engineering-physics courses, very severe difficulty, and they endeavor to avoid the ratio reasoning (especially that part associated with area) by trying to substitute in formulas without doing the qualitative reasoning as to what makes for a larger, and what for a smaller, effect in what ratio. Without practice, they fail signally in the ratio reasoning associated with f -numbers of lenses.

Chapter 10

Early Modern Physics

10.1 INTRODUCTION

There is an understandable desire in the physics community for earlier introduction of students to aspects of modern physics. In some quarters these dreams of accelerated learning extend to advocacy of injection of the results of quantum mechanics, nuclear physics, and high energy physics as early as freshman, or even high school, level. In light of what we have been learning in recent years about cognitive development and concept formation, I doubt that genuine learning and understanding of such material is feasible at such early stages. One would only cultivate blind memorization of end results to be used in artificial homework exercises and to be tested for as what Eric Rogers used to call “cheap recall.” Knowledge and understanding do *not* reside in strings of names such as “quark,” “gluon,” “neutrino,” “charm,” or “wave function.” When the “How do we know . . .? Why do we believe . . .?” questions are not being dealt with, no genuine learning or understanding can be achieved. I doubt that it is wise for us to succumb to subject matter pressure (as so many chemists have done, for example) and force our students to memorize end results without understanding.

What seems to me to be feasible and highly desirable in an introductory course is to get to the insights gained in early 20th century physics: Electrons, photons, nuclei, atomic structure, and (perhaps) the first qualitative aspects of relativity. To achieve this, it is impossible to include all the conventional topics of introductory physics. One must leave gaps, however painful this may seem. How does one decide what is to be left out? One powerful way, in my experience, is to define what I call a “story line.” If one wishes, say, to get to the Bohr atom, one should identify the fundamental concepts and subject matter from mechanics, electricity, and magnetism that will make understandable the experiments and reasoning that defined the electron, the atomic nucleus, and the photon. The selected story line would develop the necessary underpinnings and would leave out those topics not essential to understanding the climax. For students continuing in physics, the gaps would have to be recognized, ac-

cepted, kept in mind by the faculty, and closed in subsequent courses. (If the students were given a chance really to learn, understand, and absorb the most basic concepts, they would subsequently close at least some of the *seeming* gaps on their own.) Some efficiency could be gained by putting certain topics (e.g., elementary dc circuits, geometrical optics) entirely in the laboratory and not devoting them appreciable class time. Such topics are far more effectively developed in a “hands on” context in any case (cf. Sects. 7.4-7.9 and Sects. 9.17-9.18).

If one has carefully thought out the story line to be developed, it is possible to inject, along the way through earlier material, many questions and exercises that prepare the students for thinking and reasoning that will be encountered toward the end. Quite a few textbooks are now attempting to do this, but most of them include so much material that such preparatory exercises get lost in the clutter. Most textbooks that do deal with the early 20th century developments tend to go through the material so rapidly that much opportunity for physical insight and reasoning is extruded; it is the end results that are dwelt on and not the reasoning that yielded them.

It must be kept in mind that all of what we call “modern physics” deals with levels of insight not directly accessible to our senses. Students need time to absorb and comprehend the inferences drawn from the classical experiments that led to the deep insights we now assert so quickly and casually. Furthermore, the classical experiments and the reasoning they entail provide an exceedingly rich and valuable opportunity for the kind of spiralling back that has been advocated throughout this book. Many students begin to show their first reasonably firm mastery of basic concepts such as velocity, acceleration, force, mass, momentum, energy, centripetal force, electric charge, electric field strength, and magnetic B-field when they synthesize them in rich contexts such as those provided by the Thomson experiment and the Bohr atom.

In the light of the issues outlined above, this chapter will concentrate on the intellectual growth students can achieve in the study of early 20th century physics. This happens to be an instance in which at least parts of the historical development (not all the intimate details) are deeply conducive to learning, understanding, and the cultivation of some degree of scientific literacy. Unfortunately, neglect of some of these historical aspects greatly diminishes the effectiveness of many treatments of this area of subject matter. I wish to support these contentions in the following discussion.

10.2 HISTORICAL PRELIMINARIES

The insights we usually associate with the term “modern physics” began with the qualitative study of gaseous discharge and cathode rays in the 1870s and 1880s and rose to something of a crescendo with Roentgen’s discovery of x-rays in December 1895, Becquerel’s discovery of radioactivity in early 1896, and Thomson’s experiment identifying the electron in 1896-97.

Replicas of the tubes that Crookes used in his study of cathode rays are available in most physics preparation rooms, and the demonstrations are invariably of great interest to the students. There is an unfortunate tendency, however, for lecturers to rush through the demonstrations, asserting very quickly what each one implies about the cathode beam. The impact of these demonstrations can be greatly enhanced if time is allowed for contemplation, discussion, and inference. As the demonstration is performed, it is more effective to ask the students what is to be inferred from the observations. For one thing, this helps many students who are still in need of exercise in sharpening their discrimination between observation and inference;¹ for another, it allows articulation of alternative explanations and inferences—which should be tolerated and debated rather than dictatorially suppressed. Such discussion provides a very important underpinning for the study of the Thomson experiment since Thomson was motivated to resolve the debate as to whether the cathode beam consisted of particles, as had been conjectured by Crookes, or of some hitherto unknown radiation, as was being argued by Lenard (see Section 10.3). Students inclined to support a radiative model should be, at least temporarily, encouraged to do so; they would be in good company.

Many textbooks give adequate and sufficient, albeit abbreviated, discussions of the discoveries of x-rays and radioactivity. (Although there is much good physics to be learned in considering these stories in greater detail, there are limits to the time one can devote.) A few aspects, important for subsequent study, are, however, insufficiently emphasized, and students tend to lose sight of their significance. One of these aspects is the fact that both x-rays and radioactive emanations were quickly discovered to ionize air, the conductivity being observed and recognized through the discharge of electroscopes. Becquerel, in fact, initially surmised that the rays from uranium were weak x-rays. Thomson was studying x-ray induced conductivity in gases just before undertaking his classic study of the cathode beam, and his awareness of the ionization played a very important role in making the cathode beam experiments possible.

The conceptual importance of the discovery of ionization of gases is underlined by Millikan (1917):

. . . up to this time the only type of ionization known was that observed in solution, and here it is always some compound molecule like sodium chloride which splits up spontaneously into a positively charged sodium ion and a negatively charged chlorine ion. But the ionization produced in gases by x-rays was of a wholly different

¹Many students are exceedingly weak on such discrimination. Unless guided by questioning, they do not ask themselves what were the facts and evidence on the one hand and the inferences drawn from the facts on the other. They fail to make similar discriminations in other disciplines (history, for example). Yet such discrimination underlies, together with other processes, the intellectual behavior one would characterize as “critical thinking.” (See Sect. 2.19 and Chapter 13 for additional discussion.)

sort, for it was observable in pure gases like nitrogen and oxygen, or even in monatomic gases like argon and helium.² Plainly, then, the neutral atom even of a monatomic substance must possess minute electrical charges as constituents. Here was the first direct evidence (1) that an atom is a complex structure, and (2) that electrical charges enter into its makeup. With this discovery, due directly to the use of the new agency, x-rays, the atom as an ultimate, indivisible thing was gone, and the era of the constituents of the atom began. . . .

A second aspect, either not mentioned at all or too quickly glossed over, is the discovery by the Curies that a vial of radium compounds maintains itself permanently above room temperature and that, when placed in a calorimeter, "each gram of radium gives off 80 calories per hour . . . sufficient heat . . . to melt its own weight of ice." Thus a serious question was raised, from the very beginning, about the origin of all this energy and the validity of the energy conservation law.

Furthermore, both α and β radiations were shown to have mass. How could one account for the continuous emission of material particles in the apparent absence of chemical change (the work of Rutherford and Soddy on the transformation of the elements was still to come) or other alteration in the state of the radioactive material? How, in particular, could one account for material emission from elements (pure metallic uranium and radium) without interaction with atoms of other substances? Thus the law of conservation of mass was also being called into question.³

As part of awareness of their own intellectual history, it is desirable that students face and appreciate these initial questions and that the story that unfolds eventually show how they were explicitly resolved. It is through such experience that scientific literacy is enhanced, not through glossing over of the questions and rapid assertion of names and end results.

A question, which frequently arises when one elects to use elements of the historical sequence in teaching an introduction to modern physics, concerns what was known about (what we now call) Avogadro's number N_0 and about the sizes of atoms and molecules *prior* to Millikan's determination of the quantum of electrical charge and the advent of x-ray diffraction. (This is a question I am asked from time to time by my own physics colleagues.) The facts are as follows.

²It was, of course, well known that flames rendered gases conducting, but flames involved chemical reactions, introduced new products into an initially pure gas, caused convection currents and extraneous effects due to temperature differences. A steady, controlled, reproducible electrical process could not be achieved under these circumstances.

³It might be noted that it was during this period of convulsion in physical science that Henry Adams (1918) made (in *The Education*) his oft quoted remark "Chaos is the law of Nature; order is the dream of Man."

The orders of magnitude of these quantities were firmly established well before the end of the 19th century and were used in guiding both experimental work and theoretical analysis, but the values were far from precise. The sources of information were the kinetic theory of gases on the one hand and experimental data on the transport phenomena (viscosity, thermal conductivity, diffusivity) and on departures from ideal gas behavior on the other. The story began with the theoretical foundations laid by Clausius in 1857 and 1858 and by Maxwell in 1860 [see Brush (1965) for translations and reprints.] These works established the mean free path and its connection to the transport coefficients. In modern notation, for example:

$$\lambda = \frac{1}{\sqrt{2} \pi n \sigma^2} \quad (10.2.1)$$

and

$$\eta = \frac{1}{3} n m \lambda \bar{v} \quad (10.2.2)$$

where λ denotes the mean free path; n the number of atoms or molecules per unit volume; σ the atomic or molecular diameter (assuming spherical shape); η the coefficient of viscosity; m the mass of one atom or molecule; and \bar{v} the mean atomic or molecular velocity. From the Maxwellian distribution, \bar{v} is given by

$$\bar{v} = \sqrt{\frac{8RT}{\pi M}} \quad (10.2.3)$$

where R is the universal gas constant; T the absolute temperature; and M the relative atomic or molecular mass. In this notation

$$m = \frac{M}{N_o} \quad (10.2.4)$$

Combining Eqs. 10.2.1 and 10.2.2 gives

$$\eta = \frac{m \bar{v}}{3\sqrt{2} \pi \sigma^2} \quad (10.2.5)$$

Equation 10.2.5 implicitly relates the experimentally measurable quantity η to the two unknowns N_o and σ . It also contains the prediction, since n has dropped out, that the viscosity coefficient of an ideal gas is independent of the pressure—an intuitively unanticipated prediction, the confirmation of which helped provide powerful reinforcement for the newborn theory.

In 1865, Loschmidt made what appears to be the first calculation of molecular size by taking the intrinsic volume excluded by the molecules in the gas to be equal to the volume occupied by the substance in the solid or liquid state, that is, the very low compressibility of liquids and solids justifies the

assumption that the atoms or molecules are exceedingly close together in these states. If ρ denotes the density of the solid or liquid, the volume of one mole of molecules M/ρ is given by

$$\frac{M}{\rho} = \frac{1}{6} N_o \pi \sigma^3 \quad (10.2.6)$$

Combining Eqs. 10.2.5 and 10.2.6, Loschmidt obtained a value of molecular diameter. He could readily have obtained the value of n , which has come to be called the “Loschmidt number,” but did not actually do so. One can also calculate what came to be called Avogadro’s number, N_o .

After van der Waals, in 1873, put forth his modified equation of state for departure from ideal gas behavior,

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT$$

recognition that the constant b must be approximately equal to four times the volume excluded by one mole of molecules in the gas phase made possible an improved calculation based on modification of Eq. 10.2.6:

$$\frac{b}{4} = \frac{1}{6} N_o \pi \sigma^3 \quad (10.2.7)$$

If one takes modern values for nitrogen, for example, of $\eta = 178$ micropoise at 27°C and $b = 0.03913$ liters per mole (l/mol), one obtains, from the van der Waals approach, combining Eqs. 10.2.5 and 10.2.7: $\sigma = 3.1$ Angstroms and $N_o = 5.1 \times 10^{23}$.

The experimental values were considerably less accurate in the 19th century, but the preceding calculation illustrates that the orders of magnitude were right and provided reliable guidance. That the estimates were deemed important is indicated by the fact that figures such as Stoney, Lothar Meyer, and Kelvin participated in their development. More accurate determinations did not come until those of Perrin (in 1908), based on Einstein’s 1905 paper on Brownian motion, combined with experimental observation of gravitational stratification and Brownian motion of particles in colloidal suspension. Further refinement of the value of N_o awaited Millikan’s (1909, 1911) determination of the corpuscle of charge and the advent of x-ray diffraction.

Unfortunately, this story does not lend itself to use in an introductory physics course, the kinetic theory base being far beyond what is realistic at that level. I include the story here, not to advocate its use in teaching, but because it might enhance the perspective of others, as it did my own, when I first explored it. One can tell students about it qualitatively if one wishes to do so and ask them to take the assertions on faith. (Although I see strong objections to asking students to take assertions on faith in early portions of the course, when, because of past inexperience, they only feebly discriminate what is fully substantiated and what is not, I see no objection to doing so occasionally after their ability to discriminate has been strengthened.)

10.3 PRELUDE TO THOMSON'S RESEARCH

Prior to the time at which Thomson embarked on his investigation of the cathode beam, two divergent views had evolved regarding its nature: British scientists adhered to the particle model advocated by Crookes, while continental Europeans, led by Philipp Lenard (then at Bonn and later at Heidelberg), preferred a wave or radiation model.

The latter view was by no means naive. Lenard had conducted numerous careful experiments on the transmission of cathode rays through very thin metal foil "windows" in the end of the cathode ray tube and on the penetration of the rays through different gases after passing through the foil. Finding that the rays penetrated the foil and continued on for another centimeter or two, still in straight lines, Lenard became convinced that the cathode rays could not be corpuscular or material in character, but must be wave disturbances in the ether. He could not conceive material charged particles penetrating a substance as dense as the foil without deflection, and, although still numerically crude, kinetic theory was far enough advanced to make it convincing that an atomic or molecular beam would not penetrate so far in air at atmospheric pressure.

The opposing points of view of Lenard and Crookes exemplify the sharp distinction between corpuscular and wave phenomena that had emerged in 19th century scientific thought. It was believed that these two manifestations were mutually, absolutely exclusive, that any given phenomenon must be either of one class or the other, that no manifestation could exhibit both corpuscular and wavelike aspects. This complete dichotomy is something well worth leading students to think about and examine as a prelude to subsequent introduction of modern views of wave-particle duality. It sets in perspective a significant episode in intellectual history.

Another experimental fact that, at the time, stood in the way of the corpuscular hypothesis was failure to achieve electrostatic deflection of the cathode beam by passing it between capacitor plates built into the tube. It was well known that a magnetic field caused deflection of the beam in the direction that would be expected of moving negatively charged particles, but attempts to produce electrostatic deflection yielded null results. (This was one of the first obstacles that Thomson proceeded to resolve.)

This difficulty led to some questioning of a result that had been obtained by Perrin. Perrin had inserted an electrometer cup at the end of the tube opposite the cathode and had collected negative charge, as exhibited by the electroscope to which the cup was connected. The question was raised as to whether the charge being thus collected was indissolubly connected with the cathode beam or was an independent manifestation.

It was at this juncture, late in 1896, that Thomson embarked on the famous experiments that led to the determination of the charge-to-mass ratio in the cathode beam.

10.4 THOMSON'S EXPERIMENTS

Thomson (1897) set out to resolve the argument concerning the nature of cathode rays. Referring to the conflicting corpuscular and wave hypotheses, he revealed some of the factors that moulded his thought:

The electrified particle theory has, for purposes of research, a great advantage over the aetherial theory, since it is definite and its consequences can be predicted; with the aetherial theory it is impossible to predict what will happen under any given circumstances, as on this theory we are dealing with hitherto unobserved phenomena in the aether, of whose laws we are ignorant. The following experiments were made to test some of the consequences of the electrified particle theory.

Thomson first repeated Perrin's electrometer cup experiment (see Section 10.3), but, instead of placing the cup at the end of the tube opposite the cathode, he sealed it into the side of the tube where it did not directly receive the undeflected cathode beam. When the tube was turned on, the electrometer showed no charge, but when the beam was deflected magnetically so that it entered the cup, the electrometer collected negative charge. Thomson writes:

This experiment shows that however we twist and deflect the cathode rays by magnetic forces, the negative electrification follows the same path as the rays, and that this negative electrification is indissolubly connected with the cathode rays.

Although the logic may seem obvious to us, it turns out that quite a few students have difficulty seeing why Thomson took the trouble to perform this experiment even though he knew of Perrin's results, and they have difficulty articulating its significance. Few have had the opportunity to think through such a sequence in previous study; this is not the nature of conventional end-of-chapter exercises.

Thomson then attacked the problem of electrostatic deflection:

An objection very generally urged against the view that the cathode rays are negatively electrified particles is that hitherto no deflexion of the rays has been observed under a small electrostatic force . . . Hertz made the rays travel between two parallel plates of metal placed inside the discharge tube, but found that they were not deflected when the plates were connected with a battery of storage cells; on repeating this experiment I at first got the same result, but subsequent experiments showed that the absence of deflexion is due to the conductivity conferred on the rarified gas by the cathode rays. On measuring this conductivity it was found that it diminished very

rapidly as the exhaustion increased; it seemed then that on trying Hertz's experiment at very high exhaustions there might be a chance of detecting the deflexion of the cathode rays by an electrostatic force.

As a result of his previous year and a half of experimenting and thinking about conductivity in gases induced by x-rays, Thomson was very sensitive to the possible role of this phenomenon. He realized that cathode rays as well as x-rays induce conductivity, and he was well prepared to visualize the consequences. As it happened, newly developed vacuum techniques made it possible for him to test his ideas by achieving sufficiently high vacuum to suppress the conductivity:

At high exhaustions the rays were deflected when the two aluminium plates were connected with a battery of small storage cells; the rays were depressed when the upper cell was connected with the negative pole of the battery, the lower with the positive, and raised [when the connections were reversed.] The deflexion was proportional to the difference of potential between the plates, and I could detect the deflexion when the potential difference was as small as two volts.

It was only when the vacuum was a good one that the deflexion took place, but that the absence of deflexion is due to the conductivity of the medium is shown by what takes place when the vacuum has just arrived at the stage at which the deflexion begins. At this stage there is a deflexion of the rays when the plates are first connected with the terminals of the battery, but if this connection is maintained the patch of fluorescence gradually creeps back to its undeflected position. This is just what would happen if the space between the plates were a conductor, though a very bad one, for then the positive and negative ions between the plates would slowly diffuse until the positive plate became coated with negative ions, the negative plate with positive ones; thus the electric intensity between the plates would vanish and the cathode rays be free from electrostatic force . . .

As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of electricity carried by particles of matter.

Now that vacuum tubes are things of the past, students no longer encounter the concept of space charge or have occasion to visualize the behavior of ions in a rarified gas. I have, on a qualifying examination, given graduate

students all the information about the initial failure to achieve electrostatic deflection of the cathode beam, told them that Thomson achieved deflection on sufficient improvement of the vacuum, hinted that the difficulty had to do with ionization of the residual gas, and asked for an explanation, with rough pictures of what had been happening. Very few graduate students were able to give a reasonably competent answer. I suggest that qualitative physical thinking of this kind should not be eliminated from introductory physics. A rich context, such as the research being described, is an invaluable opportunity to give such thinking and visualization substance and meaning.

Thomson then went on to conduct the measurements that are very cursorily described in most textbooks: Determining the deflection on application of a single field; eliminating the unknown velocity of the particles by restoring the beam to initial position through application of crossed electric and magnetic fields; evaluating the charge-to-mass ratio of the hypothetical particles.

Much physics is left out of the majority of text presentations. For example, students are not asked to consider the significance of the fact that the beam remains coherent (the spot on the screen does not smear out) when it is deflected, either electrically or magnetically, from the initial position; it does not occur to them that there is physical information in what does *not* happen as well as in what does, and it is necessary to lead them if they are to think about such matters. The coherence, of course, supports the hypothesis that the entities in the beam are identical in their properties and in their velocity and that the equation for the trajectory of one particle applies to all.⁴

Most textbook presentations simply eliminate the velocity of the particles from the two available equations, as though it were of no interest, and solve for the charge-to-mass ratio. This is not what Thomson does. He calculates the numerical value of the velocity, shows it to be of the order of one tenth the velocity of light, and argues this to be strong evidence against Lenard's hypothesis of electromagnetic disturbance in the ether. This is physics that learners should be led to confront.

An aspect that Thomson does not dwell on, since it is trivial to a physicist, is, however, not trivial for students. I have pointed out to students that there is no observable gravitational deflection of the cathode beam whereas there is large electric and magnetic deflection, and asked how they account for the unobservability of the gravitational deflection. Many students (even graduate students on qualifying examinations) respond that the gravitational deflection is so small because the mass of the particles is so small. They do not immediately invoke what they supposedly learned about the uniformity

⁴It is worth noting that a beam of *positive* ions, formed, by acceleration through suitable electrodes, out of a region of ionized gas, *does* smear out when deflected magnetically since the beam is not homogeneous in velocity of the ions. This is why a velocity selector is necessary in a mass spectrometer. I find many graduate students completely oblivious to matters of this kind since they have never had a chance to think about the phenomena. Surely the groundwork should be laid in the introductory courses.

of g for all objects, and they do not associate the smallness of the drop with the enormous velocity. This opportunity to spiral back to earlier macroscopic concepts in an entirely new microscopic context is invaluable for assisting in mastery of the earlier concepts.

Using the method of the crossed electric and magnetic fields, Thomson then went on to determine the charge-to-mass ratio (actually he gives the mass-to-charge ratio in the old cgs electrostatic units) in tubes with electrodes of different metals (aluminum, platinum, iron) and with different residual gases (air, hydrogen, carbon dioxide). His results (quite inaccurate by modern standards) fell into a relatively narrow range—given the large room for error and uncertainty in this early undertaking.

10.5 THOMSON'S INFERENCES

In his paper, Thomson remarks on certain systematic errors that he believed made his values of charge-to-mass ratio somewhat low. At this juncture, however, he was not striving for high accuracy; rather, he was interested in orders of magnitude, and he was trying to establish whether or not the charge-to-mass ratio associated with the cathode rays varied over a large range, as it was known to do with different ions in electrolysis and in conducting gases.

The results of all his different measurements fell (converted to modern units) between 0.67 and 0.9×10^{11} C/kg. This being a very much narrower range than that observed for different ions in electrolysis, and also being within the range of uncertainty of his experimental measurements, Thomson was led to conclude that the negatively charged particles have the same charge-to-mass ratio in all cathode beams regardless of electrode material and ambient gas. He also conjectured that, when it would become possible to determine the two properties separately, it would very likely be found that the particles all have the same charge and the same mass. On this account he denoted the mass of the particles by m and the charge by the special symbol e (rather than a more conventional symbol for an arbitrary quantity of charge), and the symbol e/m is the one used to this day.

Furthermore, the additional homogeneity of the beam, reflected in the facts (1) that the spot does not smear out under electrostatic deflection (showing that, since e/mv_x^2 must be the same for all particles, they have all fallen through the same accelerating potential difference) and (2) that the velocities of the particles are all the same (as indicated by crossed field determination), implied that the particles must all originate in the immediate neighborhood of the cathode—either by ejection from the cathode or through formation by ionization of gas at the cathode. (It is now recognized that the particles are produced through bombardment of the cathode by positive ions, formed in the gas of the very imperfect vacuum and accelerated toward that electrode. The potential difference imposed in tubes of that variety was not high enough to eject electrons from the metal through field emission.)

To establish the physical significance of the observed charge-to-mass ratio, Thomson appealed, for comparison, directly to the well-established electrolytic data. For hydrogen ions, for example

$$\frac{q_H}{m_H} = \frac{96,500}{0.0010} = 9.6 \times 10^7 \text{ C/kg}$$

whereas the value for oxygen is $1.2 \times 10^7 \text{ C/kg}$. The charge-to-mass ratio of ions produced by irradiation of gases was, at that time, much less precisely established but was known to be of the same order of magnitude.

Although this may sound trivial to some teachers, many students have not acquired, in their course of study, explicit realization of the fact that numerical magnitudes convey almost no information when standing alone and that information and inference come primarily from comparison of magnitudes. The present context forms a vivid and valuable illustration. Concerning the charge-to-mass ratios, Thomson remarked:

Thus for the carriers of electricity in the cathode rays [e/m is very large] compared to its value in electrolysis. The [size of e/m] may be due to the smallness of m or the largeness of e , or to a combination of these two. That the carriers of the charge in cathode rays are small compared with ordinary molecules is shown, I think, by Lenard's result as to the rate at which the brightness of the [fluorescence] produced by these rays diminishes with the length of path traveled by the ray.

As was pointed out in Section 10.2, the orders of magnitude of mean free path and of atomic-molecular size were well established at this time, and Thomson was making use of this information in his interpretation of the data. Since mean free paths were known to be of the order of 1000 Angstroms at atmospheric pressure, and since, therefore, an atomic beam would have been scattered beyond detectability after traveling only ten times that distance, Thomson interpreted Lenard's data, showing penetration of the order of a centimeter or two, as indicating the smallness of the mass of particles in the cathode beam. This smallness implied, of course, the existence of a subatomic entity.

Students should be led to think about why one would not be inclined to ascribe the difference between the e/m values in the cathode beam and in electrolysis to a difference in e rather than to a difference in m , or even, perhaps, to differences in both properties. (Thomson, in the quotation given above, acknowledges all the possibilities but rejects the latter two without much discussion.) This is, again, a kind of thinking that students have rarely had the opportunity to engage. Everything has been presented to them in the form of polished and seemingly inevitable end results. When they begin to see the role played by faith in simplicity and order in nature, by the plausible reasoning and inductive guesswork based on such faith, they advance to a more

realistic grasp of the nature of scientific thought than the one most of them hold in the absence of this intellectual experience. Here is one of the elements of scientific literacy.

The coda to this story is provided by a few episodes that followed in quick succession. Zeeman, working under H. A. Lorentz at the same time Thomson was performing the preceding experiments, discovered the splitting of spectral lines in a magnetic field. Lorentz, with his deep grasp of Maxwell's electromagnetic theory, guided Zeeman to the classical interpretation, and they estimated the e/m associated with the line splitting to be of the order of 10^{11} C/kg. This indicated entities with this charge-to-mass ratio to be bound within the structure of atoms and not just materializing in a free state in the cathode beam.

In 1899 Thomson published a paper in which he showed that the electrical charge ejected from metals on incidence of ultraviolet light (the photoelectric effect) was associated with entities having the same e/m as the cathode rays. His son, G. P. Thomson, remarked in a lecture in 1956:

He also showed in the same paper that the negative particles emitted from a hot wire had approximately the same e/m . This really completed the proof. Opposition to the idea of particles smaller than atoms did indeed continue, but it was merely the spasmodic dying kicks of the older physics, a matter of muscular contraction rather than brain.

(The published version of the talk [Thomson, G. P. (1956)] gives more detail and uses somewhat different words.)

10.6 HOMEWORK ASSIGNMENT ON THE THOMSON EXPERIMENT

If, before starting study of the Thomson experiment, one asks students to write a one-page note on what they see to be the meaning of the word "electron" and how it is that we come to know about such entities, the resulting documents form a sobering study in their own right. The great majority of students pour forth what amounts to pure gibberish. They have all heard the term "electron" from early days of schooling, but the term never acquired anything like the meaning it has for science. There is only an imprecise half-remembered jargon, the residue of names and end results implanted without understanding. Only a very few students have the security and self-confidence to say that they had never understood the jargon and that they have no idea of what the term really means or how knowledge of such an entity originates.

Section 10.13 at the end of this chapter details a written homework assignment based on a Socratic sequence that I have used for many years in connection with the previously outlined study of the Thomson experiment. It

leads the students to address the questions of “How do we know about . . . ? Why do we believe in . . . ? What is the evidence for . . . ?” electrons. It also takes full advantage of the opportunity to spiral back to the use of physics concepts developed earlier in the course. As has been pointed out repeatedly, it is such spiralling back in increasingly rich context that helps in attaining mastery of the basic material, a mastery attained by only a very few students on the first encounter.

The assignment, although it follows Thomson’s thought and exposition quite faithfully, departs from Thomson’s paper in certain details. Thomson does not enlarge on the negligible gravitational effect on the cathode beam, and he shortcuts the derivation of the expression for the deflection of the beam under the influence of a single field with idealizations and approximations that students would find difficult to follow. The Socratic sequence has been structured to evoke explicit consideration of elements that Thomson legitimately assumed his professional readers did not need.

My experience with this assignment has been quite favorable. Some students initially consider it an unwanted burden (they have become accustomed to solutions of end-of-chapter problems without any verbalization whatsoever), but the great majority, as they find themselves progressing and grasping the synthesis, reflect on it as a very valuable and helpful learning experience. For many, the most striking aspect is the interconnection they naively find among the concepts studied throughout the course, especially the ones studied much earlier and then put aside without further use. This naivete is, in itself, a measure of the importance of such spiralling back, especially when it can be placed in a sufficiently simple, understandable, yet conceptually powerful and important, context.

10.7 THE CORPUSCLE OF ELECTRICAL CHARGE

Many 19th century scientists saw Faraday’s law of electrolysis as implying the atomicity of electrical charge. A typical statement is that of Helmholtz in a Faraday lecture at the Royal Institution in 1881:

Now the most startling result of Faraday’s law [of electrolysis] is perhaps this: If we accept the hypothesis that the elementary substances are composed of atoms, we cannot avoid concluding that electricity also, positive as well as negative, is divided into elementary portions which behave like atoms of electricity.

Since the order of magnitude of Avogadro’s number was known (Section 10.2), it was clear that the order of magnitude of the elementary charge was 10^{-19} C, obtained by dividing the Faraday (10^5 C) by Avogadro’s number (10^{24}). Attempts at direct measurement were made by Townsend through study of total charge carried on clouds of water droplets in gases irradiated

with x-rays. Thomson refined Townsend's technique and reported values between 1.8 and 2.8×10^{-19} C. As is well known, the definitive measurements were achieved with Millikan's oil drop experiment, the results of which were first published in 1911, with refinements continuing for a number of years.

The Millikan experiment is adequately treated in many textbooks, and there is no need of detailed elaboration here. The principal element lacking in many presentations, however, is the opportunity for the student to see some actual data exhibiting discreteness through the experimental scatter. There is also significance in the identification of the least common multiple among observed quantities of charge and changes of charge. [Holton (1978) gives a pedagogically useful analysis of Millikan's approach to his data.]

One verbal aspect deserves care and attention: It has become conventional in many textbooks and lectures to announce that "Millikan measured the charge on the electron." The term "electron" has had a long and complex history, and there is little to be gained in exploring this history in an introductory course. It is true that, in early discourse, the term was used in reference to a corpuscle of electrical charge regardless of its carrier, but, in modern terminology (which crystallized in the early 1900s), the word "electron" denotes the particle in the cathode beam and in the structure of atoms. This is the way students understand the term. Thus, given the modern terminology, saying that "Millikan measured the charge on the electron" becomes profoundly misleading; students get the impression that Millikan dealt directly with electrons.

What Millikan did was measure the size of the elementary charge as it was to be observed, accreted on oil droplets, in an ionized gas. Although some of the ions might have been electrons, most were not. The connection to the cathode ray particles was, of course, immediate, but it must be remembered that this was a matter of guesswork and plausible reasoning, based on faith in simplicity in nature (how would nature manage to maintain perfect electrical neutrality if the elementary charges were not all identical?) and not on direct work with electrons as such. This is a beautiful illustration of how real scientific thinking works and progresses—another aspect of enhancing scientific literacy. The opportunity is lost, however, in the misstatement that "Millikan measured the charge on the electron."

10.8 FROM THOMSON'S ELECTRON TO THE BOHR ATOM

If one wishes to bring the modern physics story at least to the point of the early quantum picture of atomic structure, there are certain ingredients—in addition to the two discussed above—necessary to the generation of a sequence in which the student can see coherence, plausibility, and intelligibility rather than just disconnected assertions of end results. Here it is not possible to

follow the historical sequence in rigorous detail. The time demanded would be excessive, and much of the conceptual and mathematical material lies at a level far beyond that of an introductory course. It is quite possible, however, to form a sequence that addresses the “How do we know . . . ? Why do we believe . . . ?” questions by using appropriate segments of the historical development; that maintains plausibility and continuity; that capitalizes on opportunities to spiral back to earlier concepts; and that makes comprehensible the necessity of the departures from classical theory. [As one example of such a sequence based principally on historical material, see Arons (1965). There are many other presentations that depart from the historical base completely but still provide a coherent and intelligible development, e.g., *PSSC Physics*.]

Among the blocks of subject matter essential for such a development are: (1) Bright line spectra of gases; their assumed connection to absorption and emission of light by accelerated charged entities on the microscopic level; and the empirically obtained Balmer-Rydberg formulae for the spectral series of atomic hydrogen. (2) Nature and properties of radioactive emanations. (3) Determination of atomic dimensions and emergence of the nuclear model. (4) The photoelectric effect and the photon concept. (Broader insight and synthesis can, of course, be achieved if items such as the role of radioactive decay in transmutation of elements, positive rays and isotopes, and properties of x-rays were to be included. There is not likely to be time for such inclusion, however, in an introductory course, and the omissions do not destroy the coherence of the story line that has been selected.)

The photoelectric effect will be discussed separately in the following section (Section 10.9). The remaining, essential items listed above are presented well in many sources and will not be discussed in detail here except for the following few peripheral comments.

1 In connection with bright line spectra: It is very effective to point out the parallelism between the role of Kepler’s empirical laws in the evolution of the theory of gravitation and the role of the empirical Balmer-Rydberg formulae in the evolution of the theory of atomic structure. Many students do not yet have a clear idea of what is meant by “empirical,” and these two episodes set the term in clear contrast with “theoretical.” Furthermore, these episodes illustrate how modern science sometimes advances through combination of Baconian empiricism with the formation of new concepts and theories and how the two modes fruitfully interact with each other. Here is still another step toward scientific literacy.

2 In connection with the nature and properties of radioactive emanations: Since α -particles were crucial to development of the nuclear model of atomic structure, it is important to set them in an intelligible perspective. Rutherford (and others), showed, through measurement of electric and magnetic deflection, that β -rays had the same e/m as the particles in the cathode rays, but Rutherford at first thought both α - and γ -rays to be “undeviable.”

Then, becoming suspicious of this conclusion for alphas, he turned to a power company for the use of a very much more powerful magnet than was previously available to him and succeeded in deviating the α -rays. After machining better pole pieces for the magnet, he measured both q/m and the velocities of the α -particles from radium and from "radium emanation" (radon) by electric and magnetic deflection. The observations were repeated later with greater accuracy [see Rutherford (1903) and (1906)]. The velocities were observed to lie between 2 and 3×10^7 m/s. Since the observed value of q/m applies to either singly ionized hydrogen molecules or doubly ionized helium atoms, Rutherford faced the problem of making positive identification.

The Curies had reported the heat generated in radium to be about 80 calories per gram per hour (cal/g/h). Rutherford had determined (by observation of scintillations on a fluorescent screen) the number of alphas emitted per unit mass of radium per unit time, and he had the velocities from the electric and magnetic deflections. From these data, he could estimate the total kinetic energy of the alphas on either assumption as to their identity, kinetic energy being twice as great if they were helium atoms than if they were hydrogen molecules. The data were crude, but agreement with the Curies' value was better if the particles were assumed to be helium.

Ramsey and Soddy had previously noted that, as radon gas decays in a sealed tube, the spectrum of helium in the residual gases becomes more intense. This led Rutherford to a definitive experiment [Rutherford and Royds (1909)]. A schematic diagram of the apparatus is shown in Fig. 10.8.1.

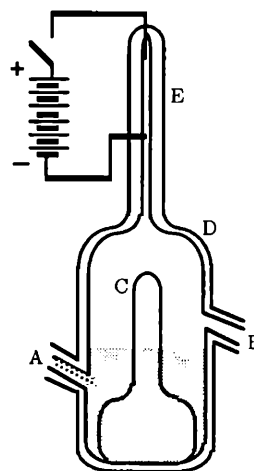


Figure 10.8.1 Schematic diagram of the Rutherford-Royds apparatus for showing spectroscopically that helium gas is formed from α -particle emanation. A mercury inlet; B vacuum pump connection; C thin walled tube containing radon gas; D heavy walled tube confining α -particles that penetrate wall of C; E heavy walled capillary into which collected gas is compressed and subjected to electric discharge.

Radon gas, obtained by pumping it away as it emanates from a radium compound, is placed in the inner, extremely thin-walled (0.001 cm) glass tube. This tube is sealed within a larger heavy-walled evacuated tube. The α -particles from the decaying radon have enough energy to penetrate the thin-walled tube but are trapped in the thick-walled outer container over a seal of mercury. After a sufficient period of time (about a week), enough gas

accumulates in the outer container to allow a spectroscopic test. Additional mercury is let in to compress the small amount of gas into the capillary, which is fitted with electrodes. Electrical discharge between the electrodes causes emission of a line spectrum from the trapped gas. Rutherford and Royds reported positive identification of the helium line spectrum and thereby settled the question as to the identity of α -particles.

(This summary is included here because this beautiful experiment is not usually described in introductory texts, yet it answers a very basic “How do we know. . . ?” question, exhibits fine, readily understandable, experimental technique, and offers an opportunity for simple, qualitative physical reasoning rarely available to the students. Thinking about experimental techniques, design, and reasoning need not—and *should* not—be limited to the introductory laboratory; it should be included in *all* parts of an introductory physics course.)

A second important aspect of dealing with radioactivity resides in exposure of the students to the arithmetic of exponential decay. Since the arithmetic is similar, this is also a valuable opportunity to make the connection with exponential growth; otherwise few students become explicitly aware of the intimate connection between the two processes.

If one is dealing with students at the calculus-physics level and has access to the exponential function, the discussion can be carried out in such terms. In this context, it is desirable to refer to the basic differential equation, and to point explicitly to other physical situations the students may have encountered (or will encounter) in which the same differential equation applies (e.g., capacitive or inductive circuit elements, decay of gas pressure due to leakage through a small hole in a container, pressure variation with height in an isothermal atmosphere, monomolecular chemical reaction, etc.).

It is not essential, however, to be able to deal with the differential equation and the exponential function. The same insights can be gained through the arithmetic of half-life or doubling time. (In fact, even the calculus-physics students need exercise in this arithmetic just as much as the noncalculus students.) Since such exercises involve ratio reasoning, many students have serious difficulty. The difficulty is greatly reduced for these students (and their security with ratio reasoning is enhanced) if they are shown how to use a simple, concrete graphical aid such as that sketched in Fig. 10.8.2 to keep track of the effect of passage of an integral number of half-lives or doubling times.

Such coupling of growth and decay calculations offers an excellent opportunity to acquaint students with Bartlett’s (1976-1979) powerful series of articles on the dangers inherent in unrestrained exponential growth.

3 In connection with atomic dimensions and the nuclear model: As soon as the size of the corpuscle of electrical charge e is available from the Millikan experiment, students can be led to calculate Avogadro’s number N_0 from the Faraday of electricity: $N_0 = 96,500/e$. Having Avogadro’s num-

ber, they can be led to calculate the order of magnitude of atomic-molecular size by turning to the basic assumption first put to such use by Loschmidt (see Section 10.2), namely that the very low compressibility of solids and liquids (compared to that of gases) implies that the constituent particles are essentially contiguous to each other. One obtains a very reasonable order of magnitude, 3 Angstroms for example, if one calculates the volume of one water molecule in liquid water ($18/N_0$) and takes the effective molecular diameter to be the side of a cube having this volume. (Students are usually handed the end results of these simple calculations but are rarely led to carry the calculations through themselves, including the rounding off to one significant figure rather than listing all eight or ten figures emerging on the display of the hand calculator.)

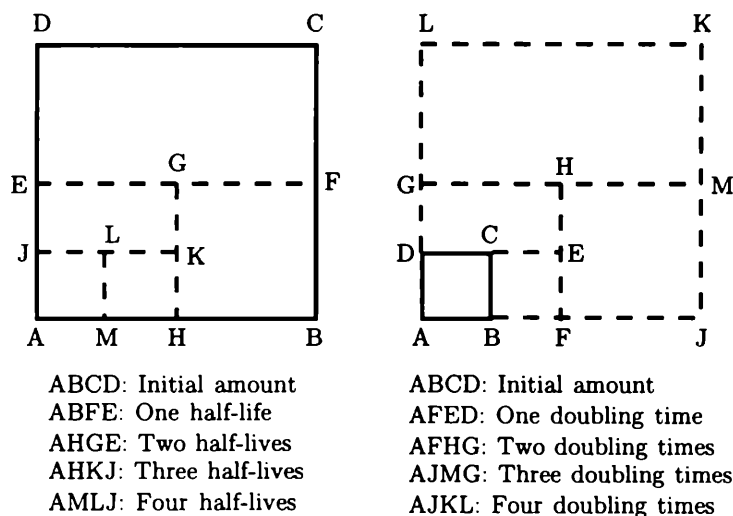


Figure 10.8.2 Keeping track of growth and decay ratios for integral numbers of half-lives or doubling times.

It is important to keep emphasizing to the students that numerical values have little or no meaning standing by themselves. Meaning is generated through comparison with other values. A valuable homework question (rarely found in textbooks) would lead students to make such comparisons after completing the above calculations: How does atomic-molecular size compare with wavelengths of ultrasound, visible light, and x-rays; with the smallest distance resolvable in a good microscope; with the “smoothness” of an “optically smooth” surface; with the average separation of molecules in a gas at atmospheric pressure?

The latter question offers an excellent opportunity to revive geometrical scaling in an intellectually significant context: Since the density of gases is roughly several thousand times lower than that of liquids and solids, the average spacing in gases must be the cube root of this ratio and therefore of the order of tens or hundreds of times atomic dimensions. Such crude but significant order of magnitude estimates are a powerful antidote to students who

have imbibed the (essentially poisonous) idea that “science is exact.” Furthermore, such experiences convey to students the insight that “estimating” in science is not wild, unsubstantiated guesswork (as most of them have come to think) but is based on careful, imaginative reasoning, however crude the numerical data may be.

The next coordinated step involves formation of the nuclear model and establishing the order of magnitude of nuclear size. The story of the backward scattering of α -particles, as observed by Geiger and Marsden and interpreted by Rutherford to imply the presence of minute but massive scattering centers, is detailed in one way or another in many texts. In relatively few texts, however, are students given the opportunity to think through the rich phenomenology that is involved and to make the simple calculation that sets an upper bound to nuclear size.

First, it is necessary to get students to go back to laboratory (or demonstration) experience with macroscopic collisions and explicitly recapture the insights: (1) that only forward “scattering” occurs when a more massive object collides with a less massive, stationary one, and (2) that bouncing back (“large-angle scattering”) occurs only when the incident object is less massive than the target. These effects may seem obvious to the teacher, but many students never really registered them on first encounter, or have lost sight of them in the interim, and do not spontaneously invoke these ideas in connection with the microscopic phenomena, undetectable to our senses, now under consideration.

A second element requires consideration of the fact that, if one tentatively adopts a nuclear model to account for the bouncing back of an α -particle, the positively charged projectile must pass through a “cloud” of negative charge in order to approach the nucleus. Rutherford, in his first paper on this interpretation, makes the full mathematical calculation for an assumed spherical distribution, but this is not really essential for the students. It is sufficient to invoke the idea (that should have been developed in connection with the inverse square laws of both gravitation and electrostatics) that the field is zero anywhere within a uniform spherical shell and that the effect of the negatively charged cloud would therefore decrease very rapidly as the α -particle penetrated it.

The simple numerical calculation the students can then make is that of the closest approach of the α -particle to the center of the target. Given the mass of the α and the velocity of the order of 2×10^7 m/s, one has the kinetic energy of the projectile, the distance of closest approach being determined by the radial separation from the target at which all this kinetic energy is stored as potential energy within the target-projectile system. (In the Geiger-Marsden observations, the target was a gold film, and Rutherford took the number of elementary positive charges on the gold nucleus to be about $100e$, or about half the relative atomic mass of gold. At this time the meaning of the atomic number in the periodic table had not yet been appreciated.) On this

calculation, the distance of closest approach is about 3×10^{-4} Angstrom—about 1/10,000 the atomic-molecular dimension. (It is important to lead the students to articulate the perception that this calculation gives an upper bound to nuclear size, not the nuclear size itself. Again we have an illustration of the making of a profoundly important *estimate* rather than obtaining an “exact” result.)

The preceding calculation invites spiralling back and can be tied directly to the students’ recapitulation of calculating, given the initial kinetic energy, how high a stone goes when thrown vertically upward, and of calculating escape velocity from the earth (or moon or planet) if the escape velocity idea has been previously invoked. The calculations should be accompanied by descriptions, in the students’ own words, of the energy transformations taking place (as previously suggested in Sections 5.3).

10.9 THE PHOTOELECTRIC EFFECT AND THE PHOTON CONCEPT

It is not possible to tell the story of black-body radiation in an intellectually honest and meaningful way in an introductory course, and hand waving about it only leaves the students mystified. The clear and intelligible way to the quantum concept is through the photoelectric effect, and this is the path almost universally adopted in current texts. Although some texts give presentations that honestly address the “How do we know . . . ?” questions, many, unfortunately, abbreviate the story to the point at which students are presented with end results that are memorized without understanding.

The bulk of the relevant experimental work is that of Philipp Lenard (1902). To understand what transpired it is necessary to know how the experiments were conducted, and it is thus essential to start with at least a schematic diagram of the apparatus (such as Fig. 10.9.1) and an insight into what was being measured. (Quite a few texts give good descriptions of what Lenard did, and only a very concise summary will be given here for the sake of illuminating the pedagogical discussion.)

Referring to Fig. 10.9.1: Electrode P is connected through a sensitive galvanometer (or microammeter) A to the midpoint G of the slide wire resistor CD. (Lenard actually made his measurements with two separate electrometers, connected to electrodes M and P, respectively.) If slide contact S is at point G, the potential difference between M and P is zero. If S is to the left of G, plate M is positive relative to P, and electrons ejected from M would be attracted back toward this electrode and retarded in their motion toward P. In the following, the convention is adopted of denoting this as a “retarding” or negative potential difference. Similarly, with S to the right of point G, ejected electrons would be repelled from M and accelerated toward P. This is denoted as an “accelerating” or positive potential difference. Light from an

arc is incident on metallic plate M through window Q. Monochromatic light is obtained by using filter F.

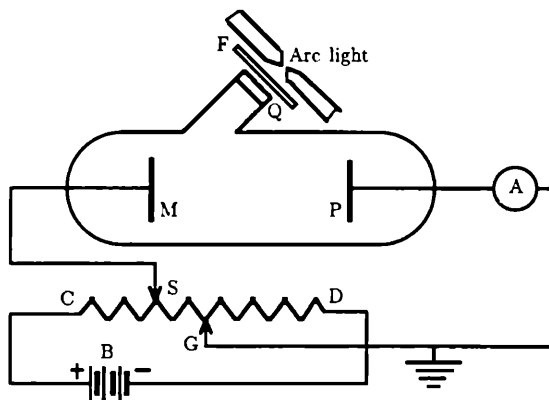


Figure 10.9.1 Schematic diagram of Lenard's apparatus for measuring photoelectric current.

Since students in introductory courses get very little experience in interpreting the functions of a simple electrical circuit, the apparatus of Fig. 10.9.1 offers a valuable pedagogical opportunity. One can use this rich context to supplement conventional numerical exercises on the photoelectric effect with the opportunity to think through what the apparatus does and how the ejected electrons are affected at various positions of the contact S. Teachers who have not invoked such situations will be shocked by the inability of many students (even those in engineering-physics courses) to interpret what happens to the potential difference between plates M and P as contact S is moved along the wire. (This is not a matter of complexity of the ideas; it is merely a matter of lack of practice on the part of the students.)

Using three different light sources (an arc light with carbon electrodes, another with zinc electrodes, and a spark discharge between zinc spheres), Lenard made a systematic study of the influence of light intensity and of the potential difference between plates M and P on the photocurrent. The leading observations are summarized as follows:

- 1 With a steady light source and a fixed, positive (accelerating) potential difference, the photo-current was observed to increase as plate P was moved toward plate M. When P came to within about 5 mm of M, the current was observed to level off at a final, maximum value. All the further observations were made at this final spacing between the plates. (Students should be led to interpret the point and purpose of these observations and the inference to be drawn: The experiment showed that electrons must be ejected at all angles relative to M and virtually all of the ejected electrons were being collected on P when the spacing was sufficiently small.)

2 Lenard then varied the light intensity in two ways: By changing the current through the arc and by moving the source to a greater distance from the window Q. (In the latter case, he made use of the inverse square law for variation of intensity with distance from a small source.) He found that the maximum (saturation) photocurrent under accelerating potential difference was directly proportional to the intensity (energy/unit area/second) of the incident light (Fig. 10.9.2).

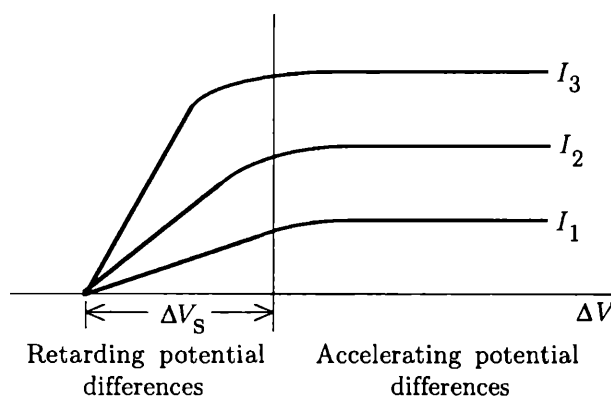


Figure 10.9.2 Idealized curves representing results of Lenard's observations of photoelectric current with apparatus such as that in Fig. 10.9.1. Curves such as those indicated were obtained with three different intensities of the same wavelength of monochromatic light. Values of saturation current (I_1 , I_2 , I_3) are directly proportional to incident light intensity. Stopping potential difference (ΔV_s) is independent of light intensity; it depends only on the wavelength of the incident light and on the material composing plate M.

3 Lenard was especially impressed by the fact that the direct proportionality to incident light intensity extended all the way down to extremely low intensities—one three-millionth of the highest intensity available—with no evidence at all of a threshold level. (A threshold was expected on the supposition that some finite level of energy flux would be required to eject any electrons at all from confinement within the metal. No such intensity threshold has ever been detected for the photoelectric effect; if the incident light is capable of producing the effect at all, some electrons are always ejected, no matter how faint the light.)

4 When contact S was moved to the left of point G (Fig. 10.9.1), and the potential difference acquired retarding values, the observed photocurrent did not drop to zero immediately despite the retarding effect. It decreased more or less linearly (Fig. 10.9.2) with increasing retarding potential difference, reaching zero at a potential difference of the order of two volts. The value ΔV_s at which the photo-current is cut off is called the "stopping potential difference." Lenard observed that the value of the stopping potential difference was

completely unaffected by the intensity of the incident light (Lenard varied the intensity of the incident light by a factor of more than 1000, with everything else held constant, without detecting significant change in ΔV_s); it was altered only when a different type of light source was used or when the metal in plate M was changed. (This was, of course, one of the most surprising aspects of the observations. It was anticipated that electrons would be ejected from the metal with a distribution of kinetic energies, but it was also expected that this distribution would be influenced by the intensity of the incident light.)

In the latter observations, one encounters again the reasoning concerning kinetic energies of ejected electrons and the storing of potential energy when a retarding field is imposed between capacitor plates. It is inevitably disappointing to a conscientious teacher, but one must steel oneself to the fact that quite a few students will still have difficulty with these concepts in the new context despite what one might have done on preceding occasions. It takes several encounters, in altered context, for many students to master these abstract concepts—not just one or two encounters. Thus the added opportunity is very useful.

Given the summary of Lenard's observations, students should be led to address questions such as the following:

- (a) Sketch several possible trajectories of electrons ejected from plate M when plate P is a centimeter or two away and when it is within 5 mm. Do this for the cases of both accelerating and retarding potential difference. Describe cases of projectile motion that correspond to sketches you have made for the behavior of the electrons.
- (b) The quantity ΔV_s is said to be equal to the maximum kinetic energy of electrons ejected from plate M. Explain the reasoning behind this statement in your own words. What is the corresponding expression for the case in which a stone is thrown vertically upward? Why is it that the mass of the particle is present in the expression in the gravitational case and absent in the electrical case?
- (c) Why is it surprising that the maximum kinetic energy of the ejected electrons is unaffected by the intensity of the incident light? Redraw Fig. 10.9.2 so as to show a set of three curves that would *not* have been surprising to Lenard. [Note that this question leads the student to confront, explicitly, what is *not* the case in contrast to what *is*.]
- (d) It is clear that, although the intensity of the incident light has no effect on the kinetic energy of individual electrons, it does affect the saturation current, the latter being directly proportional to the intensity, with no evidence of a threshold. From the variation of the saturation current, what are you forced to conclude concerning the sole effect of brightness of incident light in the process being observed? Can you, in terms of

the physics you have learned so far, account for the observed fact that the intensity of the incident light determines only the rate of ejection of electrons without having any effect on their individual energies? [Lenard explicitly points out that it is very difficult, if not impossible, to explain the observed photoelectric phenomena in terms of classical Maxwellian theory.]

Einstein, in his famous 1905 paper proposing the photon concept [see translation by Arons and Peppard (1965)], deals with the photoelectric effect only as a secondary matter. His main object is to re-derive the black-body spectrum, not as Planck had approached it through quantization of the energy of the atomic or molecular oscillators in the walls of the cavity, but through statistical thermodynamics and quantization of the radiation itself. (This material is far beyond the level of an introductory course.) Having generated the concept, however, and having successfully derived the black body distribution, Einstein then turns to the photoelectric observations for further support. He shows that the photon concept is capable of providing a very simple resolution of the photoelectric paradoxes, and these arguments are properly given in most introductory textbooks.

The essential ingredient is to give students the opportunity to present the arguments in their own words instead of desperately trying to remember textbook assertions without ever having stated these for themselves. I have seen presentations in which students are given the photon model ex-cathedra and are then asked to use it to explain the observations. This is “backwards science;” it undermines and destroys the “How do we know . . . ?” approach—thus inhibiting, rather than cultivating, scientific literacy.

The most important ingredient for future use is, of course, the linear relation between stopping potential difference and the frequency of incident monochromatic light, with its different threshold frequency for each different metal. It is important to note that this linear relation was not among the original photoelectric observations, although some textbooks imply that it was. Einstein *predicted* the linear relation, together with its now familiar interpretation, and this remained to be confirmed. Initial support was provided by Richardson and Compton (1912) and by Hughes (1913). The final and most definitive confirmation was provided by Millikan through the magnificent experiment that involved preparing a clean, uncontaminated metal surface by enclosing what he described as a small “machine shop in a vacuum” [Millikan (1916)].

Millikan subsequently said of his own work [Millikan (1949)]:

[Einstein's explanation of 1905] ignored and indeed seemed to contradict all the manifold facts of interference and thus [seemed] to be a straight return to the corpuscular theory of light which had been completely abandoned since the time of Young and Fresnel. .

. . . I spent 10 years of my life testing the 1905 equation of Einstein's, and, contrary to all my expectations, I was compelled in 1915 to assert its unambiguous experimental verification in spite of all its unreasonableness since it seemed to violate everything we knew about the interference of light.

As a step toward enhancing scientific literacy, one might note in this connection that there are other similar episodes of theoretical prediction and subsequent experimental confirmation in the history of science. Dalton, for example, predicted the law of multiple proportions before he found the regularity already present in the composition data—data that had never been converted from percentages to the form that revealed the small whole number ratios required by a corpuscular model (see Chapter 12 for more detail).

10.10 EINSTEIN'S PAPER ON THE PHOTON CONCEPT

Because of their clarity and eloquence, it is very useful to have at hand some of Einstein's own words concerning the photon concept. A few especially useful passages are quoted in the following, all being taken from Einstein (1905a) [translation by Arons and Peppard (1965)].

It should first be noted that the title of the paper is "Concerning a Heuristic Point of View Toward the Emission and Transformation of Light." Thus Einstein emphasizes the heuristic nature of his proposal. The paper begins in a highly characteristic manner; just as he does in the relativity paper, Einstein points to a certain asymmetry and lack of conceptual consistency in existing theories. The first paragraph of the following quotation points to the fact that Planck's theory of black-body radiation requires quantization of the energies of the material particles in the walls of the cavity while the electromagnetic radiation that is being emitted and absorbed is treated as continuous:

. . . the total energy of a ponderable body must, according to the present conceptions of physicists, be represented as a sum carried over the energies of the atoms and electrons [that make up the body]. The energy of a ponderable body cannot be subdivided into arbitrarily small parts [i.e., Planck's theory], while the energy of a beam of light from a point source (according to Maxwellian theory of light or, more generally, according to any wave theory) is continuously spread over an ever increasing volume.

The wave theory of light, which operates with continuous spatial functions, has worked well in the representation of purely optical phenomena and will probably never be replaced by another theory. It should be kept in mind, however, that the optical observations refer to time averages rather than instantaneous values. In spite of the complete experimental confirmation of the theory as applied

to diffraction, reflection, refraction, dispersion, etc., it is still conceivable that the theory of light which operates with continuous spatial functions may lead to contradictions with experience when it is applied to the phenomena of emission and transformation of light [i.e., interactions on the microscopic scale].

It seems to me that the observations associated with black body radiation, fluorescence, the photoelectric effect, and other related phenomena . . . are more readily understood if one assumes that the energy of light is discontinuously distributed in space. In accordance with the assumption to be considered here, the energy of a light ray spreading out from a point is not continuously distributed over an increasing space, but consists of a finite number of energy quanta which are localized at points in space, which move without dividing, and which can only be produced and absorbed as complete units. [The term "photon" emerged later.]

In the preceding paragraph, Einstein points to fluorescence as one of the phenomena that Maxwellian theory had not dealt with successfully. He is referring to what is known as "Stokes's Rule," the then well established fact that, when a substance fluoresces under incident radiation such as ultraviolet light, the "transformed" radiation (emitted light) invariably has a longer wavelength or lower frequency than the exciting radiation. Concerning Stokes's Rule, he suggests that perhaps each incident photon, absorbed by the fluorescent material, stimulates the emission of one or more photons, leading to the reemission of the energy that was absorbed. If each absorption and corresponding reemission is an elementary process, independent of other incident photons, conservation of energy requires that the energies of the emitted photons (and therefore their frequencies) be equal to or less than that of the incident photon.

Einstein then continues with regard to Lenard's photoelectric observations:

The usual conception, that the energy of light is continuously distributed over the space through which it propagates, encounters very serious difficulties when one attempts to explain the photoelectric phenomena, as has been pointed out by Lenard in his pioneering paper. According to the concept that the incident light consists of energy quanta of magnitude $h\nu$, however, one can conceive of the ejection of electrons by light in the following way. Energy quanta penetrate into the surface layer of the body, and their energy is transformed, at least in part, into kinetic energy of electrons. The simplest way to imagine this is that a light quantum delivers its entire energy to a single electron; we shall assume that this is what happens. . . . An electron to which kinetic energy has been imparted within the body will have lost some of this energy by the time it reaches the surface. Furthermore, we shall assume that in leav-

ing the surface of the body each electron must perform an amount of work W_0 , characteristic of the substance of which the body is composed. The ejected electrons leaving the body with the largest normal velocity will be those that were directly at the surface. The kinetic energy of such electrons is given by⁵

$$KE_{\max} = h\nu - W_0 \quad (10.10.1)$$

If the emitting body is charged to a positive potential difference relative to a neighboring conductor, and if ΔV_s represents the potential difference which just stops the photoelectric current, [it follows that $e\Delta V_s$ must be equal to the maximum kinetic energy of the ejected electrons] and therefore

$$e\Delta V_s = h\nu - W_0 \quad (10.10.2)$$

where e denotes the electronic charge.

If the deduced formula is correct, a graph of ΔV_s versus the frequency of the incident light must be a straight line with a slope that is independent of the nature of the emitting substance. [This constitutes the prediction of what came to be known as the Einstein equation for the photoelectric effect.]

So far as I can see, there is no contradiction between these conceptions and the properties of the photoelectric effect observed by Lenard. If each energy quantum of the incident light, independently of everything else, delivers all its energy to a single electron, then the velocity distribution of the ejected electrons will be independent of the intensity of the incident light; on the other hand, the number of electrons leaving the body will, if other conditions are kept constant, be proportional to the intensity of the incident light. . . [This covers Lenard's observed anomalies.]

10.11 BOHR'S FIRST QUANTUM MODEL OF ATOMIC HYDROGEN

Like Thomson's research on the cathode beam and Lenard's investigation of the photoelectric effect, Bohr's first simple model of the hydrogen atom, from which he obtained the Balmer-Rydberg formula for the spectral series of atomic hydrogen by making tentative departures from Newtonian and Maxwellian theory, carries an intellectual experience accessible to students in introductory physics. If carefully motivated and interpreted, at a pace slow

⁵Einstein's symbols have been altered to correspond to the notation adopted in this book.

enough to allow comprehension, the effect on many students is dramatic. They recognize that they are putting together in one context a large volume of the most fundamental material they previously encountered in bits and pieces: Circular motion, centripetal force, Coulomb's law, kinetic and potential energies, absorption and emission of light, conservation of energy, bright line spectra, electrons, the nuclear model, the photon concept. They find the synthesis to be a revelation of "interconnectedness" that they are able to comprehend. They find reinforcement in their ability to put it together, and they sense how the opportunity to spiral back gives them an increasingly firm grasp of the basic concepts. At the same time, they take a first step toward new ideas such as the correspondence principle, discrete energy levels, ground states, and excited states, all of which remain basic to the final, correct quantum theory.

There is very great pedagogical value in treating the hydrogen atom at the level at which this is done in Bohr's very first paper of 1913. Unfortunately, many textbook versions eviscerate the treatment, shortcutting it as they shortcut the Thomson experiment, thus greatly reducing the physical content, the impact, and the intelligibility. The following is a condensed outline of how Bohr handled the problem in his first paper. The analysis is confined to circular orbits, and the necessary quantization rule is obtained not through arbitrary quantization of angular momentum (such treatments are nothing but black magic to students at an introductory level) but through application of the correspondence principle, which is far more plausible and intelligible to the students at this stage even though algebraically more complex. (Section 10.14 contains an example of a Socratic homework assignment that leads students to put the story together as a unified sequence.)

Bohr begins with explicit recognition of the fact that Rutherford's nuclear model "seems to be necessary in order to account for the results of the experiments on large angle scattering of α -rays." He points to the resulting conflict with classical theory which requires an orbiting electron to radiate continuously:

. . . the electron will approach the nucleus, describing orbits of smaller dimensions, and with greater and greater frequency, the electron on the average gaining kinetic energy at the same time the whole system loses energy.⁶ The process will go on until the di-

⁶Students should have been exposed to analysis of these kinetic and potential energy changes in connection with satellite motion under gravity, and perhaps orbital motion (without radiation) of a charged particle under Coulomb's law, earlier in the course, before any mention of the hydrogen atom. Now they spiral back to recover these previously encountered ideas for the new context. Care should have been taken in the earlier treatments to show why it is not possible to take the zero reference level for energy at $r = 0$ and why the most convenient zero reference level is at infinite separation (the kinetic energy of orbital motion tending to zero as $r \rightarrow \infty$ and the potential energy of the system being conveniently taken as zero in the same limit). They should then have been led to see that the system must lose energy to the outside when the radial separation decreases, gain energy from the outside when the radial separation increases. They should also be led to see that, on this

mensions of the orbit are of the same order of magnitude as the dimensions of the electron or those of the nucleus. A simple calculation shows that the energy radiated out during the process will be enormously great compared with that radiated out by ordinary molecular processes.

It is obvious that the behavior of such a system will be very different from that of an atomic system occurring in nature. In the first place, the actual atoms in their permanent state seem to have absolutely fixed dimension and frequencies. Further, if we consider any molecular process, the results always seem to be that after a certain amount of energy characteristic for the systems in question is radiated out, the systems will again settle down in a stable state of equilibrium, in which the distances apart of the particles are of the same order of magnitude as before the process . . .

The way of considering a problem of this kind has, however, undergone essential alterations in recent years owing to the development of the [quantum theory of electromagnetic] radiation, and the direct affirmation of the new assumptions introduced in this theory, found by experiments on very different phenomena such as specific heats, photoelectric effect, Roentgen rays, etc. [Here Bohr is referring to the successes of the photon concept, in the years since 1905, in various areas to which Einstein and others applied it.]

The result of the discussion of these questions seems to be the general acknowledgment of the inadequacy of the classical electrodynamics in describing the behavior of systems of atomic size. Whatever alteration in the laws of motion of electrons may be, it seems necessary to introduce in the laws in question a quantity foreign to the classical electrodynamics; i.e., Planck's constant, or as it is often called, the elementary quantum of action. By introduction of this quantity the question of the stable configuration of the electrons in the atoms is essentially changed, as this constant is of such dimensions and magnitude that it, together with the mass and charge of the particles, can determine a length of the order of magnitude required.

In the last sentence Bohr is referring to arguments from dimensional analysis which were widely prevalent at the time. (Although students are usually shown the importance of checking the dimensional consistency of their own

account, given a finite radial separation, the total energy must be less than that at infinite separation, and therefore negative relative to the reference level adopted. If these aspects have not been slowly and carefully developed earlier, the negative sign that now arises in connection with the total energy of the system presents a serious roadblock; students panic before it, feel they cannot possibly understand the mathematics, and proceed to memorize without comprehension.

work in solving problems, they are rarely shown, in introductory courses, the powerful role that such seemingly crude thinking has had in inductive leaps made in scientific research. The present context provides a valuable opportunity for such exposure.) One line of argument involved angular momentum: It was clearly recognized at the time Bohr was writing that h has the dimensions of angular momentum, and, as a matter of fact, unsuccessful efforts had been made to produce nuclear models in which total angular momentum changed by integral amounts as individual electrons entered or left an orbital ring occupied by several electrons. Another dimensional line of argument (referred to in the last sentence in the quotation above) involved showing that the combination h^2/mke^2 has the dimensions of length and gives an order of magnitude corresponding to atomic dimensions (the expression is for modern SI units, m denoting the mass of the electron, and k the constant in Coulomb's law).

With this background of motivation, Bohr suggested a direct application of Einstein's photon hypothesis in the following manner:

1 Abandon classical electrodynamics to the extent of assuming that, at radii of the order of atomic dimensions, electrons can revolve in stable circular orbits ("stationary states") without radiating continuously. Retaining the law of conservation of energy, the electron-nucleus system is then assumed to gain or lose energy only when electrons are transferred from one stationary state to another. Thus, if an electron is transferred from an orbit r_1 to an orbit r_2 , the system must change energy by the amount $E(r_2) - E(r_1)$, the change corresponding to absorption or emission of energy depending on whether r_2 is greater or less than r_1 .

2 Invoking Einstein's photon concept, assume that electromagnetic radiation is absorbed or emitted only in transfer of electrons from one orbit to another, and that such absorption or emission of energy by individual electrons is associated with absorption or emission of individual quanta of energy $h\nu$ as suggested in Einstein's heuristic explanation of the photoelectric effect. This gives the relation

$$h\nu = |E(r_2) - E(r_1)| \quad (10.11.1)$$

3 Turning to the empirical Balmer-Rydberg formula for hydrogen and writing it as an expression for observed bright line frequencies ν (instead of as an expression for $1/\lambda$), one has

$$\nu = cR_H \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \quad (10.11.2)$$

Equation 10.11.2 strongly suggests, on comparison with Eq. 10.11.1, that the two terms on the right-hand side of the Balmer-Rydberg formula are referring to stationary states at two different radii. Thus Bohr writes what

corresponds to

$$h\nu = |E(r_2) - E(r_1)| = hcR_H \left| \frac{1}{n_f^2} - \frac{1}{n_i^2} \right| \quad (10.11.3)$$

4 Since only certain discrete frequencies ν are observed in the hydrogen spectrum, Eq. 10.11.3 implies that only certain discrete stationary states or energy “levels” (i.e., only certain discrete values of orbital radii r) occur in the structure of the hydrogen atom. This poses the problem of finding, in some way, an additional condition that restricts the allowed energy levels, that is, in modern terminology, a quantization rule.

5 Finally, there is the question of how to visualize the specific atom now under consideration, namely hydrogen. Bohr writes:

General evidence indicates that an atom of hydrogen consists simply of a single electron rotating round a positive nucleus of charge e . [This conclusion] is strongly supported by the fact that hydrogen, in the experiments on positive rays of Sir J. J. Thomson, is the only element which never occurs with a positive charge corresponding to the loss of more than one electron.

At this point Bohr gives a footnote citing Thomson (1912) in which Thomson, describing his work with a crude early version of a mass spectrometer (in which a beam of positive ions was separated by being subjected to *parallel* electric and magnetic fields, causing each species of ion present to fall on a parabolic track on the screen at the end of the tube) remarks that

All the elements I have examined give multiply charged atoms with the exception of hydrogen on which I have never observed more than one charge.

(Giving students this background of physical insight to savor makes the story far more interesting and effective than the bland assertion that hydrogen must consist of one electron and one proton because it is the lightest element known. Bohr saw fit to support his model with much more sophisticated evidence than this statement.)

6 It is now required to say something about the energy of a stationary state, and Bohr takes the classical quantity

$$E(r) = -\frac{ke^2}{2r} \quad (10.11.4)$$

as the total energy of the electron-proton system. Here he sees it necessary to say something about the mix of assumptions being made:

[It is assumed] that the dynamical equilibrium of the systems in the stationary states can be discussed by the help of ordinary mechanics, while the passing of the systems between different stationary states cannot be treated on that basis.

(Here, of course, are some of the crucial points on which the early quantum theory broke down. It was unable to say anything about the probability of the transitions, and, eventually, the mixture with Newtonian mechanics had to be foregone entirely. Bohr, however, was fully conscious of the tentative nature of his exploration, and he clearly put forth the dubious aspects.)

If only certain discrete energy levels are "allowed," it follows that the electron can occupy only certain discrete orbits of radius r_n with energy given by

$$E_n = -\frac{ke^2}{2r_n} \quad (10.11.5)$$

Also, since each of the two terms on the right-hand side of Eq. 10.11.3 must be a quantity of energy, it is implied that any particular energy level E_n might be related to the Rydberg constant R_H by

$$E_n = -\frac{hcR_H}{n^2} \quad (10.11.6)$$

where n is an integer and cannot be equal to zero.

7 Elimination of E_n from Eqs. 10.11.5 and 10.11.6 yields

$$r_n = n^2 \frac{ke^2}{2hcR_H} \quad (10.11.7)$$

and the remaining problem, as indicated in paragraph 4 above, is to find a quantization rule that restricts the allowed radii and makes possible the evaluation of the Rydberg constant in terms of the more fundamental universal constants.

8 In his subsequent papers, Bohr utilized the quantization of angular momentum in terms of $h/2\pi$ as described in most textbooks, but he did so only after he had shown in the first paper that the quantization obtained through application of the correspondence principle boiled down to this angular momentum quantization.

The correspondence principle was invoked by introducing the requirement that, as n and the orbit radii r_n become very large (approaching macroscopic dimensions), the frequency of the photon emitted in transitions between adjacent orbits becomes equal to the frequency of orbital motion, that is, the frequency of the radiation that would be emitted on the basis of classical Maxwellian theory.

Examining the behavior of ν as a function of n for jumps between adjacent orbits in Eq. 10.11.3:

$$\nu = \frac{E_{n+1} - E_n}{h} = cR_H \left[\frac{2n+1}{(n+1)^2 n^2} \right] \quad (10.11.8)$$

and, as n becomes large

$$\nu \rightarrow cR_H \frac{2}{n^3} \quad (10.11.9)$$

Applying Newtonian dynamics to the orbital motion of the electron (while neglecting motion of the nucleus, that is, treating this as a one-body rather than as a two-body problem, something for which correction was made subsequently), the frequency f_n of orbital motion is given by

$$f_n^2 = \frac{ke^2}{4\pi^2 m r_n^3} \quad (10.11.10)$$

The correspondence principle, requiring ν to approach f_n as n becomes large, suggests that, in light of Eq. 10.11.9 one should set

$$f_n = \frac{2cR_H}{n^3} \quad (10.11.11)$$

Combining Eqs. 10.11.7, 10.11.10, and 10.11.11 then gives the familiar results

$$R_H = \frac{2\pi^2 m (ke^2)^2}{ch^3} \quad (10.11.12)$$

$$E_n = -\frac{2\pi^2 m (ke^2)^2}{n^2 h^2} \quad (10.11.13)$$

$$r_n = n^2 \frac{h^2}{4\pi^2 m ke^2} \quad (10.11.14)$$

which, for angular momentum $L_n = 2\pi m r_n^2 f_n$, give

$$L_n = n \frac{h}{2\pi} \quad (10.11.15)$$

From then on, Bohr adopted Eq. 10.11.15 as the quantization rule.

Although the algebra involved in applying the correspondence principle is slightly more complex than that in applying Eq. 10.11.15, the physical reasoning in the former is much more plausible to the students. It is clear that Bohr himself initially felt the same way; otherwise it is unlikely he would have bothered to publish the correspondence principle approach before going on to quantization of angular momentum.

It is worth noting Rutherford's initial response to these ideas. In a letter to Bohr, dated 20 March 1913, Rutherford says [Birks (1962)]:

I have received your paper safely and read it with great interest, but I want to look it over again carefully when I have more leisure. Your ideas as to the origin of the spectrum of hydrogen are very ingenious, and seem to work out well; but the mixture of Planck's ideas with the old mechanics makes it very difficult to form a physical idea of what is the basis of it. There appears to me one grave difficulty in your hypothesis, which I have no doubt you fully realize, namely, how does an electron decide what frequency it is going to vibrate at when it passes from one stationary state to another? It seems to me that you have to assume that the electron knows beforehand where it is going to stop.

Some students begin to wonder about matters of this kind if their wondering is not suppressed by an implication that the questions are foolish. If they begin to wonder, they are in very good company indeed and are preparing to understand why the new quantum mechanics eventually had to replace the old.

There are several questions, in addition to the usual numerical calculations concerned with the diagram of energy levels, that help deepen student insight into the quantum model:

1 Some years ago, on a qualifying examination for graduate students, I inserted the following: Starting with a reproduction of the absorption and emission spectra of sodium, placed one above the other, came the question "Explain **QUALITATIVELY**, in a few words, how the quantum model of the atom accounts for the difference between these two spectra, that is, how does it come about that the emission spectrum has the same lines as the absorption spectrum but also has additional lines that the absorption spectrum does not contain?"

Fourteen students took the exam. One of the fourteen gave the straightforward response that, in absorption, electrons are elevated from the ground state to higher states and one would see only those transitions, while, in emission, electrons cascade down through intermediate states, in addition to dropping directly to the ground state, and thus produce additional lines. Five students, despite the emphasis on "qualitatively," launched into irrelevant quantum mechanical formalism with selection rules and reached no conclusion. The remaining students left the question blank. These were not incompetent students; they simply had never had the opportunity to confront basic questions of this kind and talk about them in their own words. This experience underlines the need to expose students to such qualitative questions from the earliest encounter.

2 By the time Bohr proposed the quantum model, it had been noted that the first 10 or 12 lines of the Balmer series could be observed in laboratory discharge tubes, while as many as 33 lines of this series had been detected in

stellar spectra and in the corona of the sun. Bohr turned this observation to good account:

[This] is just what we should expect from the above theory, [according to which] the diameter of the orbit of the electron in the different stationary states is proportional to n^2 . For $n = 12$ the diameter is equal to 160 Angstroms, or equal to the mean distance between molecules at a pressure of about 7 mm mercury. For $n = 33$ the diameter is equal to 1200 Angstroms, corresponding to the mean distance of molecules of about 0.02 mm mercury. According to the theory, the necessary condition for the appearance of a great number of lines is therefore a very small density of the gas; for simultaneously to obtain an intensity sufficient for observation, the space filled with gas must be very great.

The 7 mm pressure is approximately that in a laboratory discharge tube; it was difficult to obtain sufficient intensities at lower pressures. One has here an opportunity to use numerical values of the radii to achieve physical insight into a significant set of phenomena. In the discharge tube, the atoms begin to interfere with each other around $n = 12$, whereas such interference does not arise in stellar atmospheres until around $n = 33$. In the latter case, one has the benefit of being able to look at a tremendous volume of gas, and observable intensity is achieved despite the very low concentration. This is valuable physical thinking in which students cannot engage unless explicitly afforded the opportunity.

3 An opportunity to spiral back to basic electromagnetic concepts is afforded by one of the outstanding failures of the Bohr model: The prediction of a nonzero magnetic moment for the ground state of atomic hydrogen, whereas the ground state is known to have zero magnetic moment. Many students have, at this stage of the game, forgotten about current loops and the attendant magnetic behavior. Review of this concept and connecting the macro- and microscopic phenomena help register the ideas more firmly. This encounter, at the same time, paves the way for deeper appreciation of one of the early triumphs of the modern quantum theory—at least for those students who continue to that level.

Final comments: The Bohr hydrogen atom is, of course, an essentially ad hoc model and is not the result of powerful theoretical synthesis. The theory that we call “quantum mechanics” evolved later. This does not, however, make the Bohr story pedagogically useless. Much of its vocabulary and physical insight are retained to this day. Its intelligibility to students provides a rational step to modern insights. Similarly, the photon concept has evolved over time to the point that Einstein’s completely particle-like localization of the entity has been abandoned. This does not mean that the story must be abandoned (after all, Einstein’s Nobel prize was for this development rather than for relativity.)

One need not conceal from students the fact that the story continues. [Kidd, Ardini, and Anton (1989) give a discussion of the evolution of the photon concept from Einstein to the present and provide an extensive bibliography.]

10.12 INTRODUCING SPECIAL RELATIVITY

Einstein's revolutionary demolition of the classical notions of absolute space and time took a long time to penetrate the scientific community at large. Quick understanding and acceptance came only to a relatively small number of already prepared minds. The ideas involved in the transition to special relativity are exceedingly subtle and contra-intuitive. They are more subtle and abstract, even when reduced to careful operational thought experiments, than the law of inertia or the abstractions associated with energy, momentum, and electricity. It is little wonder that students, even physics majors, emerge from their first exposure with virtually no conceptual understanding of what has transpired, regardless of how well they might do end-of-chapter problems manipulating the consequences of the Lorentz transformations.

Because of the subtlety of the ideas, there is no quick and easy way of infallibly capturing all beginners. It is quite possible, however, to provide a qualitative, phenomenological introduction that lays the groundwork for subsequent better understanding of the formalism on the part of science students and also gives those who will not go further some comprehension of what is meant by the relativity of space and time. A rather abbreviated outline of how this might be done is given in the following. Readers seeking greater detail will find such treatments, for example, in Arons (1965) and Huggins (1968).

As a start, before going on to dealing with, or even mentioning, different frames of reference, it is necessary to re-examine the ways in which we measure both space and time in a single frame—the one most familiar to us. Many textbooks give adequate operational descriptions of the measurement of length, and this issue will not be belabored here. The principal difficulty, as far as student understanding is concerned, relates to what Bridgman (1962) called “spreading time over space,” that is, synchronizing clocks that are widely separated and giving meaning to “simultaneity.” Although such synchronization, the use of signals, and so on, is mentioned in many texts, the missing ingredient is usually that of sufficiently strong emphasis on the fact that spreading time over space, in our *own* frame of reference, involves *definition*, *convention*, *invention* on our part and is not already out there independent of us. This may sound trivial to a physicist long acclimated to these ideas, but, in fact, there is very great resistance to accepting this view among most newcomers—young or old. They do not believe that clock synchronization and simultaneity must be *defined* by an agreed-upon convention. (That Einstein himself regarded the concept of “simultaneity” as crucial to understanding is indicated by the prominent role he gave it in his own popularization of relativity [Einstein (1961)].)

It is necessary to separate two aspects of simultaneity: Local and distant. When we speak of simultaneity, most students immediately think only of the former—the sense we have about events taking place together, right in front of us, here and now—and they fail to see that there is a problem concerning establishing simultaneity (or non-simultaneity) for a remote event with one taking place before us, or for two remote events. Bridgman (1962) points out that we accept “local simultaneity” intuitively as a *primitive*, as mathematicians now accept “point” and “line,” without futile attempts at definitions that turn out to be circular. “Distant simultaneity,” however, requires careful operational definition, and that is where synchronization of remote clocks comes in. Novices find it difficult to accept the idea that our intuitive sense of local simultaneity does not automatically extend to remote simultaneity, and many of them strongly resist the idea that what we do with clocks is a definition and not just the quantification of pre-existing reality. (Resistance can be shaken to some extent by asking the student how we would ever establish what is happening on α -centauri, four light years away, right *now*.)

Without full awareness of the fact that synchronization is a matter of definition even in a *single* frame of reference, students are unprepared to understand what happens when we start comparing observations from two different frames. On the other hand, having developed such awareness, they are better prepared to comprehend the disagreements between different observers. Thus the introduction to special relativity is more effective if it starts with a slow, careful look at spreading time over space in a single frame and with strong emphasis on the fact that “distant simultaneity” must be operationally *defined* and is not something that already “exists” independently.

There are a variety of valid treatments and gedanken experiments concerning clock synchronization and simultaneity available in the literature, each with its own merits. For the purpose of later use in comparing assessments of simultaneity and length measurement in different frames of reference, however, it seems to me that the simplest and most direct gedanken experiment is that in which time is spread over space (along the x -axis) by synchronizing clocks placed at equal distances on either side of a central point from which a light pulse is emitted. The two clocks are defined as having been started simultaneously by arrival of the spherical wave front from the central point. (Later, it will be easy to see that, from the point of view of another observer moving along the x -axis, the two clocks could not have been started simultaneously.)

While still in our own single frame of reference, however, it is important to emphasize for the students that the use of light (radio signals, actually) is a matter of convenience and precision rather than of logical necessity. To the best of our knowledge, based on what we know of internal consistency, other, cruder procedures (e.g., sound signals in still air, slow transport of accurate chronometers) would yield synchronization in agreement with each other and with the electromagnetic signal. The reason for concentrating on the electromagnetic signal (apart from its high precision and the fact that

this is the way world clocks are actually synchronized) is that, in the final analysis, we will find that the velocity of this signal is the only velocity that two different frames of reference (i.e., frames moving relative to one another) have in common. Hence this velocity eventually provides the only way of linking observations made in the different frames.

A second aspect that merits slower discussion than is afforded in most texts is that of the passionate 19th century search for the “absolute” frame of reference, which came to be identified with the electromagnetic ether. Many students fail to comprehend the motivations for the search. They have not yet fully absorbed the role of frames of reference in physical theory; they are shaky on the meaning of “inertial frame”; and, unless explicitly prompted, they do not perceive why 19th century physicists hoped that the ether might turn out to be the primary frame for both mechanics and electrodynamics. Some discussion of this background, and the opportunity it affords for spiralling back to basic aspects of frames of reference, greatly strengthens their grasp of these underlying aspects of physical theory.

Many students, if given the chance to speak up, show themselves to be uneasy about the use of the word “absolute” in this context, not being at all certain what it means. Such students are in good company, as indicated by Bridgman’s comment:

The sort of tacit idea that we have before us in using the word “absolute” is itself not very definite, and may be one thing for the theologian and philosopher and another for the physicist. I think most physicists have in the back of their heads when using the word “absolute” not something which cannot be specified in terms of physical operations, as did the theologian and philosopher, but something in which . . . the operations can be specified in terms which do not refer to accidental, temporary, local situations. . . . Thus if the existence of an all-pervading ether could have been established in some way, then velocity with respect to the ether would have been the sort of thing that the physicist would have been willing to call absolute. It is curious that there is a uniquely definable velocity, namely velocity with respect to the fixed stars, which is not felt to have the property of absoluteness implicitly wanted. . . .

(It is interesting to speculate on what Bridgman might have said about the discovery of the 3°K cosmic microwave background radiation and our motion relative to it.)

After providing an understanding of what motivated the search for the earth’s motion through the ether, one appropriately goes on to the experimental attempts and the accompanying null results. These, especially the Michelson-Morley experiment, are well discussed in many texts. Concerning the point of departure into special relativity, Einstein’s own statement in the

second paragraph of the 1905 paper is still one of the best and clearest. After pointing out, in the first paragraph, certain unhappy asymmetries in the prevailing view of electrodynamic phenomena, he continues:

Examples of this sort, together with the unsuccessful attempts⁷ to discover any motion of the earth relative to the "light medium," indicate that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest. They suggest rather that . . . the same laws of electrodynamics and optics will be valid for all frames of reference for which the laws of mechanics hold good. We will raise this conjecture (the purport of which will hereafter be called the "Principle of Relativity") to the status of a postulate, and also introduce another postulate which is only apparently irreconcilable with the former, namely that light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body. These two postulates suffice for the attainment of a simple and consistent theory of the electrodynamics of moving bodies based on Maxwell's theory for stationary bodies. The introduction of a "luminiferous ether" will prove to be superfluous inasmuch as the view here to be developed will not require an "absolutely stationary space" provided with special properties, nor assign a velocity vector to a point in empty space in which electromagnetic processes take place.

In connection with the second postulate concerning the velocity of light, it is worth noting Bridgman's (1962) remark that, "The postulate that the velocity of light is independent of the velocity of its source is indispensable to relativity theory and is a much more fundamental postulate than that of the equality of velocity in all frames of reference. . . ."

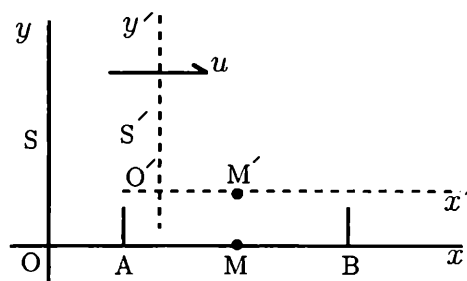
Having synchronized clocks in a single frame by means of light signals from a central point, and having motivated Einstein's postulates, it is now possible to lead students to perceive that observers in a second frame of reference, moving relative to the first, will not agree that clocks in the first have been synchronized by the chosen operation. It is also easy to see which clock the moving observer believes to have been started ahead of the other.

Referring to Fig. 10.12.1, S represents the first frame; M is the midpoint between locations A and B at which, by definition in S, simultaneous events

⁷There has been much discussion by historians of science in recent years as to whether or not Einstein was aware of the Michelson-Morley results at the time of writing the 1905 paper. Many minute details, pro and con (mostly con), have been adduced. Einstein himself in interviews, late in his life, with Shankland (1963) made contradictory statements about this. As far as an introduction to the subject and Einstein's thinking are concerned, however, it is immaterial whether he was aware of Michelson-Morley specifically. It is clear from this introductory paragraph that he was fully aware of at least some of the null results of efforts to detect motion relative to the ether.

are triggered on arrival of a light or radio pulse originating at M. S' represents the second frame, moving to the right at velocity u relative to S . The points M and M' coincide at the instant of emission of the light pulse.⁸ The first step (and this takes time and effort, quick assertion leaves a blank field) is to lead the students to comprehend that the postulate concerning the velocity of light requires that observers in S claim that the spherical pulse is permanently centered on M and that all of S' is moving to the right relative to the center of this sphere. At the same time, they must be led to see that observers in S' claim the light pulse to be permanently centered on M' while all of S is moving to the left relative to the center of the sphere. They must visualize the difference in the claims of the respective frames, and they must learn to accept these views as forever irreconcilable. This is what takes time.

Figure 10.12.1 Light flash is emitted from point M midway between remotely separated points A and B along the x -axis in frame S . By definition in frame S , arrival of the pulse synchronizes clocks or triggers simultaneous events at A and B. Frame S' moves to the right at velocity u relative to S . Points M and M' are coincident when the flash is emitted.



Having set up the basic situation as just outlined, one can proceed to inject further detail:

From the point of view of observers in S' , with M' permanently centered in the hoop (referring to the suggested demonstration equipment), points A and B have been moving to the *left* within the spreading circle. Thus, with B moving *toward* the advancing wave front and A moving *away* from it, S' claims that the signal would have arrived at B *before* arriving at A and that the clock at B was therefore started *before* the clock at A. (The reciprocal view, namely that S claims that a clock at A' in frame S' would have been started before a clock at B' , follows by the similar line of argument and makes an appropriate homework problem.)

This approach to the relativity of simultaneity is well known and widely used, and I summarize it here not because it is new but because it is the necessary basis for the qualitative description of length contraction that follows. The purely qualitative description of length contraction is also well known, but it appears in very few textbooks.

The next step in the argument involves another operational concept that

⁸In class discussion, each frame of reference can be effectively represented in concrete form by means of a long board with vertical dowels mounted at each of the three indicated positions. Each frame can carry its own color to distinguish it. An instantaneous position of the spreading circle of the wave pulse originating at M can be represented by a hoop placed so that it is centered on M.

students have never confronted and with which they initially have substantial difficulty: That of the measurement of length of a moving object. The problem here is analogous to that of distinguishing between local and remote simultaneity, but, in a sense, even more subtle. Previous experience has been only with objects stationary in our own frame of reference, and one can put a ruler on the object, or mark off the ends against an available scale, at leisure, without considering any time element at all. Thus there is no expectation that time might get inextricably mixed up with length measurement. The latter is a new and very unsettling idea that is not quickly assimilated.

It is necessary to lead students to review the operation of length measurement when they possess the rod, say, in their own frame of reference (defining “proper length” while at it) and then explicitly raise the question of what operations would be necessary to measure the length of a rod which is flying by in another frame. It takes very loaded questioning to extract the unfamiliar and unanticipated notion that one must mark the ends of the flying rod *simultaneously* against the scale in one’s own frame of reference.⁹

Once one has arrived at the perception that the ends of the moving rod must be marked simultaneously, length contraction follows directly from the previously established failure of simultaneity. Suppose that, in Fig. 10.12.1, S' is holding a rod parallel to the x' -axis; the proper length is then L'_0 . Observers in S measure the length of the moving rod by marking the ends simultaneously, according to their own clock synchronization, along the x -axis. How does S' view this operation? Since, according to S' , any clock in S was started ahead of any clock to the left of it, the observers in S must have marked the right end of the rod *before* they marked the left end. Since the rod must have moved over to the right during that time interval, the marks are too close together, and the length L measured in S must be smaller than the proper length L'_0 . (The reciprocal argument develops in exactly the same way if S is holding the rod, and the writing out of this part of the story is well left as a homework assignment.)

One has now attained some of the major insights associated with the Special Theory: (1) That clock synchronization and remote simultaneity are a matter of *definition* in a single frame of reference and not an a priori. (2) That observers in a second frame of reference, moving relative to the first, will not agree with the clock synchronization or simultaneity of events as defined in the first frame, and vice versa. (3) That the operation of measuring the length of an object gives different values in different frames. (4) That the

⁹ An alternative, and valid, operation is, of course, to measure the time it takes the rod to pass a fixed point in our frame and then to calculate its length from the time interval and the velocity. Some students come up with this idea and should be reinforced for doing so. It is not possible, however, to discern length contraction in this operation without going to the additional step of discerning time dilation. The two operations are, of course, eventually found to agree with each other, and such internal consistency helps support the theoretical structure.

measurement of length is inextricably intertwined with the spreading of time over space. (5) That the only reason we can compare the measurements made in the different frames is that they all still have something in common, namely the velocity of light.

The qualitative development summarized up to this point provides an effective beginning for any group of students coming to the concepts *de novo*. It works with high school students as well as with more mature groups. It marks a valid endpoint for a brief introduction to the revolution in point of view toward space and time that is entailed and to the meaning of “relativity” in this context. It is also an appropriate beginning for students who are going to go ahead to the development of the formalism. Very few students who are taken directly to the Lorentz transformations and the subsequent formal derivations of time dilation and length contraction develop the direct insight into the operations and the differences between observers that stem from the qualitative introduction described above. Their understanding of the formalism is significantly sounder if the qualitative introduction (with suitable homework problems) has been provided.

Following the qualitative introduction outlined above, one can start developing the attendant algebraic relations without going directly to the Lorentz transformations. The expression for time dilation follows directly by setting up a “light clock” along the y -axis and comparing the proper time interval between two events for the frame carrying the clock with the longer interval that would be calculated in the other frame. (This is done in just this way in many textbooks.) Given the expression for time dilation, one can directly derive the expression for length contraction. Given both time dilation and length contraction, one can derive the expression for the failure of synchronization (the difference S' contends exists between clocks separated by a distance x in S —clocks that S contends are synchronized).

Given the three results now listed, one can put them together to obtain the Lorentz transformations. This approach yields the transformations as a consequence of a line of physical argument that assembles the operations and calculations made in each frame. [Derivations following this line can be found in Panofsky and Phillips (1962) and in Arons (1965).] This line of argument turns out to be far more intelligible to many students than the approach of seeking “that linear transformation of coordinates which leaves the velocity of light invariant for the two frames.” To the majority of students who have not yet developed the mathematical insights of a born theoretical physicist, the latter approach has virtually no meaning; the end results are memorized, and the underlying arguments are not assimilated.

Once the Lorentz transformations are at hand, one can derive the other usual consequences of Special Relativity. There are many valid sequences and treatments in existing textbooks, and my own experience does not single out any one approach as pedagogically superior to others. The teacher should use what he or she finds most congenial.

One comment remains to be made concerning the transition from kinematics to dynamics. For many years it was conventional to enter the discussion of dynamics through derivation of the relativistic mass, that is, the mass-velocity relation, and this is probably still the dominant mode in textbooks. More recently, however, it has been increasingly recognized that relativistic mass is a troublesome and dubious concept. [See, for example, Okun (1989).] Not only does it get one into the infelicities associated with longitudinal and transverse masses, but it also tempts one to associate relativistic mass (rather than just rest mass) with gravitational effects. The latter association is basically incorrect. The sound and rigorous approach to relativistic dynamics is through direct development of that expression for momentum that ensures conservation of momentum in all frames:

$$p = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (10.12.1)$$

rather than through relativistic mass. Unfortunately, it is more difficult to derive the momentum expression in a simple way than it is to obtain the mass-velocity relation from the collision gedanken experiments prevalent in the literature. [See Peters (1986) for a recent effort to simplify this derivation.]

In some textbooks and presentations, the velocity v in the given frame of reference (as in Eq. 10.12.1) is confused with the relative velocity u of one frame with respect to another (as in the Lorentz transformations), and students tend to confuse the two velocities even when the presentation concerning the distinction is clear. It is necessary to call attention to the distinction forcefully and explicitly by extracting a statement of it from the students themselves.

10.13 HOMEWORK ON THE THOMSON EXPERIMENT

[Note to the instructor: The specific wording of such an assignment would have to be adjusted to what background students might have been given in class or lecture.]

This a written homework assignment. All you need do is follow the sequence of questions given below; they constitute the full outline. Make this a continuous story, explaining each step of your reasoning in your own words and interpreting the results.

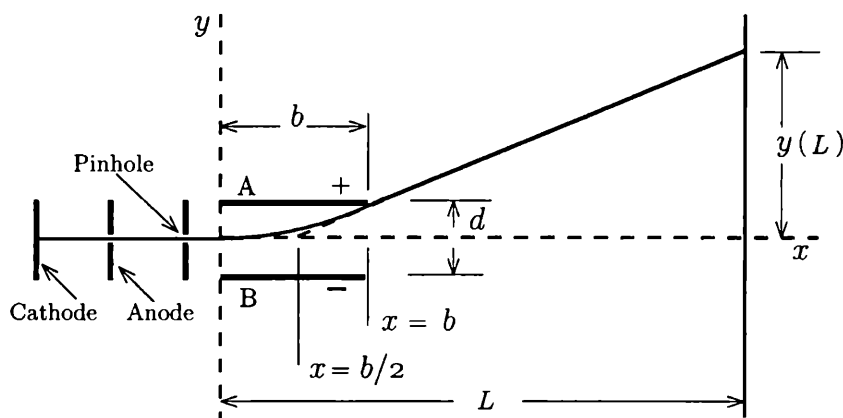
- 1 Briefly describe in your own words the goal of Thomson's investigation of the cathode rays.
- 2 Briefly describe the observations (before Thomson) indicating that cathode rays might somehow be connected to negative electrical charge. What was the point of Thomson's experiment with the charge collector placed on the side of the tube instead of at the end opposite the cathode?

As indicated in class, observations of deflection of the cathode beam (prior to Thomson) had revealed deflection by a magnetic field but had failed to reveal deflection by an electrostatic field. Thomson conjectured that the presence of ions in the

tube (the ions being charged atoms or molecules formed in the residual gas in the tube) might lead to conditions in the neighborhood of the capacitor plates such that the beam, even if it consisted of charged particles, would be unaffected by the charged capacitor plates. On greatly improving the vacuum in the tube, Thomson found that charged capacitor plates did indeed deflect the beam.

- 3 In your own words, and with accompanying pictures or sketches, explain how the behavior of numerous ions in the residual gas would create a condition such that the cathode beam would be unaffected by the charged plates. Would there still be zero electrical field *outside* the capacitor plates? Explain your reasoning.

The following questions involving algebraic equations and relationships all have to do with Fig. 10.13.1. Refer to this figure accordingly. You may redraw the figure for inclusion in your paper or you may simply include in your paper the figure that is given. All the derivations that are carried out have initially to do with the behavior of a single negatively charged particle of mass m carrying charge of magnitude e . Be sure to make this aspect clear in your presentation. After the behavior of the single particle has been analyzed, you will consider what would happen if the beam consisted of enormous numbers of particles.



Suppose that the hypothetical charged particle, starting somewhere between the cathode and anode, is accelerated to a velocity v_x by the accelerating potential difference ΔV_{acc} and enters the region between the deflecting capacitor plates (plates A and B) with this velocity. (We shall assume that the deflecting electrical field is sharply confined to the region between the plates A and B and that there are no appreciable fringing effects.)

- 4 Temporarily ignoring any gravitational effects, argue that the particle will be subjected to a uniform upward acceleration while it is between the deflecting plates A and B and will therefore follow a parabolic trajectory from $x = 0$ to $x = b$. (Note that the situation here is exactly like the one you studied in projectile motion in introductory mechanics.) What will be the character of the trajectory after the particle gets beyond the edge of the plates? Why? How is this section of the trajectory oriented (or connected) to the parabolic part?

- 5 Now derive the equations for the trajectories described verbally in part 4: Starting with the basic definitions of electrical field strength and potential difference, argue that, if we denote the potential difference between the plates of the deflecting capacitor by ΔV_{def} and the separation between the plates by d , the electrical field strength \mathcal{E} between the plates is given by

$$\mathcal{E} = \frac{\Delta V_{\text{def}}}{d} \quad (10.13.1)$$

Then, with $y(x)$ denoting the vertical deflection as a function of x , show that, while the particle is in the region between the plates, the trajectory is given by

$$y(x) = \frac{1}{2} \frac{\mathcal{E} e}{m} \frac{x^2}{v_x^2} \quad (10.13.2)$$

Then show that the slope of the parabolic trajectory at $x = b$ is equal to

$$\frac{\mathcal{E} e}{m} \frac{b}{v_x^2} \quad (10.13.3)$$

Now go back to the elementary analytic geometry of the equations of straight lines, and, making use of the point-slope form (or any other form you wish), show that the equation of the trajectory in the region between the deflecting plates and the screen is given by

$$y(x) = \frac{\mathcal{E} e}{m} \frac{b}{v_x^2} \left(x - \frac{b}{2} \right) \quad (10.13.4)$$

and use analytic geometry to prove that the straight line trajectory extrapolates back to intersect the x -axis at $x = b/2$. Letting $y(L)$ denote the vertical deflection of the particle from its initial undeflected position on the screen, show that

$$y(L) = \frac{\mathcal{E} e}{m} \frac{b}{v_x^2} \left(L - \frac{b}{2} \right) \quad (10.13.5)$$

- 6 It is an observed fact that the spot produced on the screen by the cathode beam does not smear out but remains a sharply defined spot on being deflected from its initial position when one connects a battery to the deflecting plates. If the beam does consist of charged particles, there cannot be just one; there must be enormous numbers. In the light of these observed facts and in the light of Eq. 10.13.5, what can we conclude about what the host of charged particles must have in common? Explain your reasoning.
- 7 Identify the known (measurable and observable) quantities in Eq. 10.13.5, and identify the unknown quantities. Note that there are three of the latter and that they are all properties of the hypothetical particles.

Thomson set out to reduce the number of unknowns. He hit upon an ingenious way of measuring v_x . Starting with the measured deflection $y(L)$ under known deflecting field strength \mathcal{E} , he introduced a magnetic field B (in a direction perpendicular to the plane of the figure) by means of Helmholtz coils and increased the current in the coils until the spot was returned to its initial position on the screen. Let us denote the strength of the magnetic field that brings the spot back to zero on the screen by B_0 . This is a measurable quantity.

- 8 Establish the direction of the magnetic field (in or out of the plane of the paper) that restores the beam to its original undeflected position on the screen. Going back to the basic equations for forces on charged particles in electric and magnetic fields, draw a free-body force diagram of a single particle when under the influence of the crossed fields \mathcal{E} and B_0 . Then, starting with basic concepts you studied in electricity and magnetism, show that the velocity of the particle must be given by

$$v_x = \frac{\mathcal{E}}{B_0} \quad (10.13.6)$$

Since it is an observed fact that the spot remains coherent (is not smeared out) under the influence of the magnetic field, what can you infer about the velocities of all the particles? Sketch what you might have seen on the screen if the particles did *not* all have the same velocity. Explain your reasoning. What is the significance of the fact that the billions of particles in the beam all have the same velocity? (Hint: Would they all have had the same velocity if they had originated at different points in the region between the accelerating plates? Why or why not? Since they all do have the same velocity, what is their most likely point of origin?)

- 9 Thomson calculated the velocity of the particles and found it to be of the order of one tenth the velocity of light. He then argued that the cathode beam could not be electromagnetic radiation as Lenard had contended. Put his argument in your own words.
- 10 Combining the relevant expressions that have been developed, show that, although e and m of the particles cannot be obtained separately, one can now obtain a numerical value for the *ratio* of charge to mass because it is connected with measurable quantities by the relation

$$\frac{e}{m} = \frac{\mathcal{E}}{bB_0^2} \frac{y(L)}{(y - \frac{b}{2})} \quad (10.13.7)$$

- 11 Suppose that in a given tube b and L are 4.00 cm and 20.0 cm, respectively, and that the spacing between the deflecting plates is 1.50 cm. Under a potential difference of 150 V on the deflecting plates, the deflection of the spot on the screen is observed to be 2.6 cm. The magnetic field that restores the spot to the center of the screen has a strength of 4.5×10^{-4} tesla (or webers/m²). Calculate the charge-to-mass ratio and the velocity of the particles in the beam. (Do not report more than the legitimate number of significant figures.)
- 12 Note the velocity of the particles calculated in part 11. Calculate the vertical deflection the particles sustain under the influence of gravity as they traverse the tube. Would this deflection be observable? (In answering the last question, compare the expected gravitational deflection with the order of magnitude of wavelengths of light and with the order of magnitude of the size of atoms or molecules.) Now explain in your own words why the gravitational deflection is so small and why it was proper to neglect it in any calculations made in connection with this experiment. (Be sure not to include any more significant figures in your results than are justified by the data.)

- 13 Denoting the accelerating potential between cathode and anode by ΔV_{acc} , argue that, if the negative particle originated essentially at rest at the cathode, its velocity v_x on passing through the anode would be given by

$$v_x^2 = \frac{2e\Delta V_{\text{acc}}}{m} \quad (10.13.8)$$

Because of complications with fringing fields, configuration of the cathode and anode, and because of the difficulty of measuring the high accelerating potential difference (tens of kilovolts), Thomson could not make use of this relation to obtain a reliable value of v_x even after he had evidence as to where the particles originated (see part 8). Equation 10.13.8 does, however, indicate, in a general way, what happens to v_x as the accelerating potential changes.

- 14 Suppose the accelerating potential is increased while the deflecting potential remains unchanged. What would happen to $y(L)$, that is, would the deflection on the screen increase, decrease, or remain unchanged? What would happen to the value of B_0 required to bring the beam back to $y = 0$? Explain your reasoning clearly in each instance.

Thomson made many determinations of e/m while changing conditions in the tube. He tried several different metals as cathode material (aluminum, platinum, iron). In addition to air, he made observations with other residual gases in the tube (hydrogen, carbon dioxide). (Note that even though the vacuum was quite high, there was still a significant amount of gas present.)

- 15 Why did Thomson conduct these experiments? What inferences are to be drawn from the fact that he kept getting the same value of e/m (within a fairly large experimental scatter)?

When water is decomposed by electrolysis, it is found that the passage of 96,500 C of charge liberates 1.0 g of hydrogen at the cathode and 8.0 g of oxygen at the anode. It is also well known that a molecule of water consists of two atoms of hydrogen and one of oxygen.

- 16 Calculate the charge-to-mass ratio of the hydrogen and oxygen ions in electrolysis, explaining your reasoning. Compare these values with that obtained for the particle in the cathode beam.

In connection with the values you have examined in part 16, Thomson writes: "Thus for the carriers of electricity in the cathode rays e/m is very large compared to its value in electrolysis. The size of e/m may be due to the smallness of m or the largeness of e , or a combination of these two."

- 17 State in your own words the considerations that make it seem plausible that the smallness of m best accounts for the large charge-to-mass ratio of the cathode particle rather than either of the other two possibilities cited by Thomson. (The principal observations involved are the seemingly perfect electrical neutrality of ordinary matter and the large penetration through air of the cathode beam as noted by Lenard. The latter aspect was discussed in lecture.)

- 18 In light of the entire story you have now put together, define the term "electron," that is, what does this word mean and what does it apply to?

10.14 WRITTEN HOMEWORK ON THE BOHR ATOM

This is a written homework assignment similar to the one we had earlier on the Thomson experiment. Be sure to think through everything carefully for thorough understanding. Explain lines of reasoning in your own words. Show intermediate steps of algebraic derivations. Show the numerical setup (i.e., the numerical substitution you have made in an algebraic expression) as well as the final numerical result of calculations. Pay careful attention to valid numbers of significant figures.

Much profoundly significant thinking in science is done by a process called “dimensional analysis” in which one examines combinations of physical quantities in terms of the dimensions to which the combinations reduce. Complex combinations that end up with very fundamental dimensions may (or may not) point the way to deep underlying scientific connections. Bohr starts off his epoch-making paper of 1913 with a dimensional analysis that does turn out to be profoundly significant.

- 1 Bohr introduces his paper by pointing to the fact that the combination h^2/mke^2 has the dimensions of length and that the numerical value of the combination is about 20 Angstroms. Confirm both of the preceding statements about the combination, showing your work in detail.

Note, as Bohr did, that this numerical value of length is of the general order of magnitude of atomic size (what eventually turns out to be missing is simply the numerical factor $4\pi^2$ in the denominator, see Eq. 10.14.16). Bohr argued that this is evidence of the fact that Planck’s constant h probably has some deep connection with atomic structure and should appear in any equations derived for this structure.

- 2 Other investigators, not only Bohr, had also noted that h itself has the same dimensions as angular momentum. Verify this fact. (This encouraged the view that h might have something very directly to do with angular momentum within the atom itself. We shall return to this possibility later in our analysis.)

Bohr attacked the problem of putting together a model of the hydrogen atom that would account for the observed discrete (bright line) spectra. He first had to settle on the *constituents* of the hydrogen atom, and he took it to consist of one proton (forming the nucleus) and one electron in circular orbit around the nucleus. To justify this choice of constituents, Bohr cites experimental work reported by Thomson. Thomson had investigated positive ions formed in various ionized gases (such as hydrogen, oxygen, nitrogen, carbon dioxide, ammonia, etc.) and reported that he had observed singly ionized atoms and molecules of all the substances. He also reported that, when he elevated the accelerating potential of the cathode beam that ionized the gases in the first place, he began to observe doubly ionized species. He then remarked that he had been able to obtain doubly ionized *atoms* of every species except hydrogen.

- 3 Interpret this story in your own words: How do you account for the fact that doubly ionized atoms and molecules were not produced until the accelerating potential of the ionizing beam was elevated? What was highly suggestive about the observation concerning hydrogen, that is, how does this observation support Bohr’s choice of constituents of the hydrogen atom? What other information concerning atoms also supports this choice?

Bohr then postulated that the electron occupied a circular orbit around the proton nucleus and did not radiate continuously at its orbital frequency (as required in classical electromagnetic theory) when its orbit was of microscopic scale, that is, of atomic

size. He postulated that an electron could occupy an orbit of given radius indefinitely in what he called a “stationary state.” He further postulated that the electron gained or lost energy by absorbing or emitting a photon of frequency ν , and, in doing so, “jumped” from one stationary state (orbit of radius r_1) to another stationary state (orbit of radius r_2). Since, in each stationary state, the electron-proton system has a specific total energy $E(r)$ [we shall derive the expression for $E(r)$ shortly], Bohr was saying he was assuming that

$$h\nu = |E(r_2) - E(r_1)| \quad (10.14.1)$$

- 4 Explain in your own words what lies behind Eq. 10.14.1: What motivates its introduction (i.e., what is the connection to Einstein’s heuristic picture of the photoelectric effect)? Would the energy emitted or absorbed in the jumps have any direct relation to the orbital frequency f of the electron’s motion? Why have the absolute magnitude signs been introduced in Eq. 10.14.1? Under what circumstances is the right-hand side of Eq. 10.14.1 positive and under what circumstances is it negative?
- 5 Now go back to fundamental classical physics with respect to energy quantities in orbital situations such as the one under consideration:

Show that the potential energy (P.E.) of the electron-proton system, when the electron is in a stationary state of radius r , is given by

$$\text{P.E.} = -\frac{ke^2}{r} \quad (10.14.2)$$

This involves going back to the definition of potential energy, setting up the relevant integral, choosing and justifying the choice of a zero level of potential energy, and carrying out the integration over appropriate limits with careful and correct treatment of all algebraic signs. Be sure to include an appropriate picture and force diagram. Be sure to explain your reasoning, especially in the choice of the zero level of energy, including a clear statement of why it is not possible to take $r = 0$ as the reference level.

Then find the expression for the kinetic energy (K.E.) of the electron in a stationary state of radius r in terms of the same quantities that occur on the right-hand side of Eq. 10.14.2.

Finally, show that the *total* energy $E(r)$ of the electron-proton system, relative to a reference level at infinite separation of the particles, is given by

$$E(r) = -\frac{1}{2} \frac{ke^2}{r} \quad (10.14.3)$$

Why is the right-hand side of Eq. 10.14.3 negative? How can a total energy possibly be negative? Interpret Eq. 10.14.3: Does the total energy increase or decrease when an electron is moved to a “higher” orbit (larger r)? If a photon were emitted (in accordance with Bohr’s picture), would the electron end up in a higher or a lower orbit? Explain your reasoning clearly and carefully.

Now consider the Balmer-Rydberg formula, which gives the wavelengths of lines actually observed in the various series of atomic hydrogen spectra:

$$\frac{1}{\lambda} = R \left(\frac{1}{n_2^2} - \frac{1}{n_1^2} \right) \quad (10.14.4)$$

where R stands for the number $10,973,731.2 \text{ m}^{-1}$. (The value of R is obtained *empirically*, not theoretically, i.e., it is calculated from the *measured* wavelengths. Note the extremely high precision attained in modern spectroscopic measurements!)

6 Show that Eq. 10.14.4 can be revised to yield the relation

$$h\nu = hcR \left(\frac{1}{n_2^2} - \frac{1}{n_1^2} \right) \quad (10.14.5)$$

Bohr pointed out that the right-hand side of Eq. 10.14.5 contains two separate terms and that, if one adopts the idea behind Eq. 10.14.1, each one of these terms might be interpreted as related to the total energy of a stationary state of the electron.

Putting together the observed facts of the existence of bright line spectra and the postulates of the model being developed, one arrives at the conclusion that only certain discrete stationary states (i.e., only certain special orbital radii) are allowed to exist in the electron-proton system. Present in your own words the argument that leads to this conclusion. (What would be the nature of observed spectra if all values of r were allowed?)

Argue that, if only certain special values r_n of r are allowed, only certain special energy levels E_n of the system will be possible, and show that these energy levels can be expressed in either of the following two ways:

$$E_n = -\frac{ke^2}{2r_n} \quad (10.14.6)$$

$$E_n = -\frac{hcR}{n^2} \quad (10.14.7)$$

7 Combining Eqs. 10.14.6 and 10.14.7, show that they imply that the radius r_n of the orbit associated with the integer n is given by:

$$r_n = n^2 \frac{ke^2}{2hcR} \quad (10.14.8)$$

Argue that the radius r_1 associated with the integer $n = 1$ ought to be the lowest allowed value of the radius of the electron orbit, that this should be the “normal” or “unexcited” state of the hydrogen atom, and that higher values of n and of r_n should be associated with larger orbits and “excited” states. In this context, what is meant by the term “excited”? (The $n = 1$ state is now called the “ground state” of the atom.)

The crucial problem now becomes that of finding out how nature selects or defines the “permitted” or “allowed” values of r_n from among the infinity of continuous values of r . Bohr, in his first paper, attacked this question via what he called the “correspondence principle.” This is the idea that requires any strange, new numerical behavior, on a new level of experience, to merge smoothly into what has previously been established as correct in well explored levels of experience. (In relativity for example, one applies the correspondence principle when it is required that the Lorentz transformations for position, time, and velocity reduce to the ordinary classical relations at low velocity and when it is required that the momentum and energy formulas do the same.) Bohr applied the correspondence requirement in the following way:

We know that, when oscillating on macroscopic scale (e.g., in radio antennas or in macroscopic circular orbits in magnetic fields), electrons radiate electromagnetic waves having the same frequency as the frequency of their periodic motion. Bohr therefore argues that, although at small values of r the electron does not radiate at all while in a fixed orbit (that is the real meaning of “stationary state”) and although at such radii the frequencies of the emitted or absorbed photons have no direct relation to the frequencies of orbital motion, nevertheless, as the orbits become very, very large (i.e., approach macroscopic scale), the frequency of a photon emitted in a jump between adjacent orbits should become more and more nearly equal to the frequency of the orbital motion. If that were to be the behavior, the correspondence principle would be obeyed.

- 8 Going back to the classical dynamics of circular motion, show that the frequency f_n of orbital motion of an electron in an orbit with radius r_n would be given by

$$f_n^2 = \frac{ke^2}{4\pi^2 m r_n^3} \quad (10.14.9)$$

where m denotes the mass of the electron.

We want the frequency ν of the photon emitted in jumps between adjacent orbits to approach f_n at large n . To achieve this, we need to look at what happens to the photon frequency ν as n becomes very large. Develop the following argument in detail:

Show that the frequency ν of the photon emitted in jumps between adjacent orbits is given by

$$\nu = \frac{E_{n+1} - E_n}{h} \quad (10.14.10)$$

and that (making use of Eq. 10.14.7)

$$\nu = cR \left[-\frac{1}{(n+1)^2} + \frac{1}{n^2} \right] = cR \left[\frac{(2n+1)}{(n+1)^2 n^2} \right] \quad (10.14.11)$$

Now argue that, as n becomes very large

$$\nu \rightarrow \frac{2cR}{n^3} \quad (10.14.12)$$

where the right arrow symbol means that the value on the left-hand side keeps getting closer and closer to the value on the right as n increases.

Argue in your own words that we can satisfy the correspondence principle if we introduce the requirement that

$$f_n = \frac{2cR}{n^3} \quad (10.14.13)$$

- 9 Now assemble the algebraic consequences of what has been done up to this point: You have Eqs. 10.14.7, 8, 9, and 13. Use them to solve for the quantities R , E_n and r_n in terms of the fundamental constants, that is, obtain the following relations:

$$R = \frac{2\pi^2 m (ke^2)^2}{ch^3} \quad (10.14.14)$$

$$E_n = -\frac{2\pi^2 m (ke^2)^2}{n^2 h^2} \quad (10.14.15)$$

$$r_n = n^2 \frac{h^2}{4\pi^2 m k e^2} \quad (10.14.16)$$

Finally show that these results combine to give the Balmer-Rydberg formula for lines in the atomic hydrogen spectra, with the Rydberg constant no longer simply an empirical value but fully accounted for in its relation to fundamental constants.

Discuss and interpret these results: How did the energy levels come out to be discrete? Calculate the size of the normal or unexcited hydrogen atom. Would there be any meaning to setting n equal to zero? Why or why not? Calculate the energy and wavelength of a photon that would just ionize a normal hydrogen atom. Explain your reasoning carefully.

Sketch the essence of Eq. 10.14.16 for the allowed orbits by sketching a set of at least five orbits to scale. Account for the various spectral series (Lyman, Balmer, etc.) by showing what transitions correspond to various observed lines.

- 10** Look up the meaning of “angular momentum” and write the expression for the orbital angular momentum L_n of the electron in terms of r_n and f_n . Then, making use of the expressions you have derived, show that everything reduces to

$$L_n = n \frac{h}{2\pi} \quad (10.14.17)$$

We said earlier that investigators had noted that h had dimensions of angular momentum and suspected that it might have something to do with angular momentum within the structure of the atom. Equation 10.14.17 shows that angular momentum is “quantized” on the microscopic scale. What does this mean?

In his subsequent papers, Bohr no longer used the correspondence principle to derive the results as you derived them above. He used the approach given in many textbooks and introduced “quantization of angular momentum” as one of the basic postulates instead of using Eq. 10.14.13. One obtains, of course, exactly the same final results.

- 11** Show this last statement to be correct by re-deriving Eqs. 10.14.14, 15, and 16 by using Eq. 10.14.17 as the quantization rule without invoking the correspondence principle.

Chapter 11

Miscellaneous Topics

11.1 INTRODUCING KINETIC THEORY

A few textbooks plunge directly into derivation of the pressure formula in kinetic theory without saying anything at all about the underlying assumptions; others make a few cryptic assertions concerning the model being adopted but do not try to justify the assumptions through appeal to prior experience available to the student. Only a very few textbooks discuss the assumptions in detail and provide justification.

Very few students become conscious of such gaps on their own. If no mention is made of assumptions, few realize that assumptions are being made. If the assumptions are asserted rapidly and cryptically, few students pay them any serious attention unless something along such lines is called for on tests, and then the assumptions are simply memorized with little or no consideration of how they are motivated or justified. Failure to lead students through selection of the assumptions and articulation of justifications deprives them of a rich intellectual experience with phenomenology since the underlying assumptions of kinetic theory tie to many of our everyday experiences with behavior and properties of material substances. Students need help in articulating these connections in order to understand and appreciate the structure being generated.

Furthermore, the kinetic theory of the ideal gas is an essential step in the formulation of the microscopic model underlying macroscopic properties and behavior. It is generated in the Galilean tradition of idealization and simplification. The idealizations are an essential feature. They must be fully understood, and even the existence of the computer does not obviate this necessity.

The primary feature in approaching kinetic theory is, of course, the acceptance of discreteness rather than continuity in the architecture of matter. Students are so used to having heard the terms “atoms” and “molecules” from early schooling that they are unaware that they have not examined any of the “How do we know. . . ? Why do we believe . . . ?” questions and have

been exposed only to a string of names and unsubstantiated assertions. They have no idea that other views might have been legitimately held and that, over a long period in Western thought, atomism was considered atheistic, evil, and heretical and had to be subtly defended by its relatively few courageous proponents. It would seem that cultivation of genuine scientific literacy requires at least some attention to this background. Very few textbooks, even in chemistry, any longer deal with it, however, and an individual teacher must decide for him- or herself whether or not to spend time examining the story.

Although, after having accepted discreteness either through examining evidence or by assertion, one does not wish to expend a great deal of valuable time discussing alternative models, there is something to be said for making students aware that alternative atomistic models were indeed entertained by major figures in the history of science. (A few students, in fact, think of these other possibilities but hesitate to bring them forth because of fear of being considered “stupid” if they do not immediately see the inevitability of the canonical model. Such students find significant reinforcement in knowing that they were in good company even if these initial thoughts did not survive subsequent tests.) The more important alternative models considered at various stages were the following:

- 1 *The Static Model.* Some atomists, Newton and Dalton among them, held the view that the corpuscles of a stationary, nonflowing gas occupied fixed positions and filled the entire space available to them, expanding and contracting and remaining in contact with each other as the volume occupied by the gas was increased or decreased. (Quite a few students think of this picture initially and do not see, without discussion and consideration, why the modern kinetic model must be regarded as superior.) In the *Principia*, Newton shows that if the corpuscles repelled each other with a force inversely proportional to the distance between their centers, the pressure of the gas would vary inversely as the volume, just as Boyle had demonstrated experimentally. The model therefore had the sanction of very high authority, and it persisted for a long time.
- 2 *The Boscovich Model.* In 1758 the Serbian scientist Roger Boscovich suggested a model in which matter was to be viewed as composed of indivisible point centers of force. The point centers possess inertia and interact with each other to infinite distances as do gravitating bodies. However, the force between two point centers is repulsive when they are very close together, alternates between attraction and repulsion as the points are moved farther apart, and becomes an inverse square attractive force when the points are widely separated. Thus, in a sense, Boscovichean atoms are infinite in extent. The whole conception involves an attempt to describe material substances only in terms of centers of

force and to dispense with naive notions of “stuff” and “matter.”¹ Although Boscovich’s model was not fruitful enough to achieve wide acceptance, it did influence the thinking of major figures such as William Hamilton, Michael Faraday, and Joseph Henry.

- 3 *The Vortex Model.* Early in the 19th century, Humphrey Davy proposed a qualitative dynamical theory of heat suggesting that in solids the vibration of atoms or molecules increased as the material was heated, whereas in gases the atoms rotated about their axes or possessed rotating “atmospheres.” For a brief time Joule and other investigators turned to this model and attempted to account for the tendency of gases to expand by visualizing (in an essentially Cartesian tradition) the atoms as spinning, fluid vortices, tending to expand centrifugally when external confining forces were relaxed.
- 4 *The Kinetic Model.* In his treatise on the mechanics of fluids, published in 1738, Daniel Bernoulli suggested an atomistic model based on visualizing gases to consist of minute corpuscles, moving freely and eternally at high velocities in the volume in which they are confined, exerting a steady average pressure on any boundary by virtue of extremely high frequency of bombardment. This model was neglected for about a century, one of the principal impediments being the reluctance of the scientific community to accept a model that required perfect elasticity in the microscopic interactions. The model was revived quite independently by 19th century scientists (in particular Waterston,² Maxwell, and Clausius) who were now strongly motivated to construct a dynamical theory by the advent of the concept of conservation of energy. The kinetic theory, simple and immediately enormously successful in a wide range of applications (see, for example, Section 10.2), became the basis of our modern view.

11.2 ASSUMPTIONS OF THE KINETIC THEORY OF THE IDEAL GAS

Time spent in leading students to understanding and acceptance of the basic assumptions of the kinetic model is well invested because of the range and richness of the phenomena and experiences that must be invoked. It is such hitching together of seemingly disparate, unrelated physical manifestations

¹During the 19th century two opposing philosophies were influential in science, and each had its prominent adherents. Proponents of “Naturphilosophie,” Oersted and Faraday among them, looked for the “interconvertability” of all forces in nature and accorded “force” a leading role in the conceptual structure. The positivists, on the other hand, aimed at removing from physics what they saw to be the mystique and vagueness associated with the concept of “force.”

²See Brush (1961).

that leads to mastery of concepts and understanding of the nature of scientific thought. Following is an outline of the thinking involved. The best and most effective way of conveying it to the students is to lead them to articulate the insights through Socratic group discussion rather than through didactic assertion in text or lecture.

- 1 Having accepted the atomic-molecular picture, a next step is to examine some of the immediate consequences in highly personal terms: If all matter, including our own flesh and bones, consists of discrete particles, what keeps individual particles hanging together to form liquids and solids? Our finger resists being pulled apart under tension; it also resists being compressed. If the structure is discrete, there is no alternative but to accept the existence of interactive forces between the particles. Furthermore, in liquids and solids the particles must be in “equilibrium” locations and spacings (potential wells, if the way has been prepared for the jargon) such that they attract each other very strongly if pulled farther apart and repel each other very strongly if pushed closer together.

That the interaction is probably electrical is the insight that began to develop during the 19th century with the acceptance of atomism and with simultaneous perception of the dominant role of electricity on the microscopic scale (frictional electrification, mobility of electrical charge, polarization, metallic conduction, electrolysis, ionization, dielectric breakdown). Very few students see such connections for themselves, but they begin to articulate them under questioning, and they readily appreciate their significance. Given this crude, initial insight, many students naturally want to know more about the mechanism and details of the interaction, and they expect pat and simple answers. It is a healthy experience for them to recognize that immediate jargon conveys no understanding; that at least four generations of sophisticated scientists asked the question and lived and died without arriving at an answer; that they (the students) might have to defer seeing the answers until they have learned more of the intervening physics.

- 2 Having established this initial unsophisticated insight, one can turn to gases and what might make them so different from liquids and solids. One prominent, key property is that of compressibility: Very high compressibility of gases and exceedingly low (albeit not zero) compressibility of liquids and solids.³ Given such explicit consideration of macroscopic

³This requires some thought and discussion. Many students, in the absence of previous instruction, initially believe liquids and solids to be completely incompressible since they do not experience visual evidence of compressibility. Here one must return to phenomena that transcend direct sense experience (see Section 3.12). Two commonly performed lecture demonstrations help in this context. One is the breaking under tension of a metal rod that, having been expanded on heating, cools with its ends pinned in a massive frame. The other is the breaking of a closed container of water on freezing. Students should be led to visualize

phenomena, students begin to perceive that the difference between liquids and solids on the one hand and gases on the other can be readily accounted for on the microscopic level by visualizing the discrete particles to be very close together in the former and far apart, relative to their own size, in the latter. Furthermore, they see that such a picture is consistent with the vastly lower density of gases (the order of magnitude of 1000 for the ratio of densities of solids and liquids on the one hand to gases on the other should be kept in mind) and the fluidity of gases. Having formed this picture, they can now see that, as a first approximation, it would be reasonable to take the long-range interactions of the particles to be negligible in the gaseous state and to expect strong repulsive interaction only during the short interval of direct collision. Furthermore, it becomes reasonable to expect a next approximation in which the longer range interaction would manifest itself as an attractive one (van der Waals forces) leading to condensation into liquid at lower temperature and higher pressure. With this background, students begin to see the volume of liquids and solids as representing, to a first crude approximation, the volume occupied by the atoms or molecules themselves since the low compressibility implies that the particles are virtually contiguous (see Section 10.2 for early use of this idea in arriving at an estimate of Avogadro's number). They can be led to interpret the density ratio of 1000 or more as an indication that the average separation of particles in a gas must be of the order of at least the cube root of 1000 or 10 times the size of the particles. (This is an opportunity to return, in a rich and conceptually important context, to ratio reasoning and scaling, the difficulty of which has been emphasized in earlier chapters.)

- 3 Now one can turn to evidence that the particles of the gas are in translational motion. One appropriate bit of evidence, commonly cited, is that of diffusion of odors or of colored gases (e.g., bromine) through air. A caveat, however, is posted by the confounding phenomenon of convection under small temperature differences, and more evidence is desirable. Another commonly cited aspect is the tendency of the gas to fill the entire space available to it. This is highly relevant, but it needs reinforcement by consideration of a concomitant aspect few students think of spontaneously yet perceive with just a hint. (This involves something that does *not* happen. The importance of being aware of what does *not* happen as well as of what does has been pointed to repeatedly; here is still another instance.) If one calls attention to the picture formed

that the breaking in each case occurs not because the materials are intrinsically inextensible or incompressible but because the surrounding structure is being asked to extend or compress the objects to their *initial* length or volume and that, although not impossible, such extension or compression requires *very* large forces.

up to this point (discrete particles in gases, spaced much farther apart than their own diameters) and asks what would happen in the room if the gas molecules were stationary, students will reply that the molecules would fall out onto the floor in a very thin layer, but very few perceive this without the hint. The combination of the observations listed above presents a fairly compelling basis for the initial assumption of perpetual translational motion. The assumption is subsequently reinforced by feedback from the success of the picture of pressure as stemming from collisions with the wall.

- 4 Perpetual motion of the particles immediately implies collisions with each other and with the walls of the container. Here one confronts what was the greatest impediment to early acceptance of the kinetic theory: The question of perfect elasticity of the collisions and hence of perpetual motion on the microscopic scale. Students are not as sophisticated about this problem as were mature scientists of an earlier day, but, if given the chance, they do express concern about perfect elasticity. They know that macroscopic collisions are inelastic, that motion would run down in a macroscopic system. They have been told repeatedly that “perpetual motion is impossible,” and they do not immediately perceive the possible differences at the microscopic level. First they must be led to perceive that, if one accepts the picture developed so far, there is no choice but to accept elasticity (in the average, overall behavior) since the pressure on the walls does not diminish and the gas molecules never do fall out to the bottom of the container, implying that the translational motion persists.

In fact, there are, of course, many inelastic events and interactions occurring on the microscopic scale. Individual collisions at the walls are certainly not perfectly specular. The molecules of the gas are in thermal equilibrium with electromagnetic radiation in the container, and that means continual emission and absorption of photons as quantum interactions keep occurring. As rotational and vibrational modes of the molecules are excited with increasing temperature, the frequency of inelastic collisions increases with transitions up and down in the internal modes. Although one of the most widely prevalent phenomena encountered in everyday experience, thermal equilibrium (like frictional interaction and diffuse reflection of light) is one of the most intricate and difficult to follow in full detail.

How far one is to pursue visualization of the microscopic interactions is a matter of choice for the teacher. This is more a question of time than of conceptual difficulty. There is great richness in the physics and in the opportunity to connect previously considered (and highly idealized) macroscopic phenomena with events transcending our senses on the microscopic scale.

The principal point at issue is that, in aggregate overall behavior, it is clear that microscopic inelastic events “average out” and combine into the equivalent of perfect elasticity on the macroscopic scale. This should be taken as one of the primary pieces of evidence supporting the “principle of detailed balancing”—the idea that for any given microscopic event of collision or interaction, whether elastic or inelastic, there is, instant by instant, an equal probability of exactly the reverse event in the huge macroscopic population.

- 5 Acceptance of perpetual motion on the microscopic scale leads to visualization of the intrinsic randomness of the system and the concept of distribution of speeds and directions of motion. At some point in the sequence, either here or earlier, it is, of course, highly effective to put into use one of the many available demonstrations of the kinetic model that utilize a multiplicity of small beads kept in continual random motion, or a comparable computer simulation. (In my own experience, the material beads present far more conviction to the students than do the computer simulations—at least in the early, unsophisticated stages of development. The computer simulations become useful and more impressive as one gets further along into more sophisticated aspects.) The demonstrations make vivid the fact that the particles keep changing their speeds and directions of motion, that collisions with the walls and among particles take place at all possible angles, and that instantaneous velocities range from zero to very high values and that one cannot think of a single velocity but must think in terms of distributions and averages.
- 6 It is now appropriate to visualize, without any formalism and in a purely qualitative way, the generation of pressure on the wall of the container. Here one invokes, of course, the situation that should have been examined in single particle dynamics—the force associated with change of momentum on elastic bouncing from a wall—and sets it in the context of force per unit area and an enormously high rate of collision. Discussion should elicit the perception that the force per unit area should depend both on the frequency of collisions (and thus on the density of the gas) and on the velocity of the incident particles. (This combined dependence may seem obvious to us after long familiarity with the model, but, in fact, it eluded many of the early thinkers and constitutes a sophisticated penetration on the part of the students.) Ultimately one finds that the collision frequency also depends on the velocity, and the pressure thus depends on the square of the velocity, but this must be allowed to emerge from the more formal derivation.
- 7 The model now merits further refinement through examination of the implications of the isotropy of the gas and the steadiness of pressure on the wall. How is it that the pressure remains so steady despite being

caused by individual collisions? How is it that isotropy is maintained? Students respond with relatively little difficulty to those aspects that are associated with high frequency and large numbers, that is, they see that the constancy of pressure can stem from the enormous numbers of collisions and they can visualize the fluctuations that would set in as the numbers were decreased. The moving-bead demonstrations help a great deal in this context.

What is more subtle and difficult to perceive (and what merits slower and more careful discussion) is the “principle of detailed balancing”—the notion, already mentioned in part 4 above, that constancy and isotropy must stem not simply from large numbers as such but from the fact that the large numbers lead to a situation in which any change in speed and direction of motion of a particle somewhere in the gas is invariably compensated by exactly the reverse change of another particle, always maintaining the steady state once it has been achieved from some initial, transient situation (such as the crowding of the particles toward one end of the box).

Another aspect depending on the principle of detailed balancing is the justification for treating the bouncing of molecules from the wall as, on the average, specular reflections. Students should be led to sketch a “molecule’s eye” view of the wall and realize that the wall must be “rough,” full of bumps and hills and valleys on the molecular scale. It is the averaging out of the departures from “angle of incidence equals angle of reflection” over enormous numbers of collisions that justifies the idealization of specular reflection at the walls.

- 8 Having gone this far with the qualitative picture, it is possible to anticipate that temperature will be somehow related to molecular velocity. Most students are aware that the pressure of a gas increases when the temperature is increased without change in volume (i.e., at constant density). With constant density, given the model developed so far, the only way that pressure can increase is through an increase in velocity of the molecules. Thus, temperature must be intimately connected to molecular velocity. Now the uniformity of temperature in a gas at equilibrium can also be related to the large numbers of particles and the principle of detailed balancing.
- 9 Finally, one can invoke the energy concepts and connect them with the evolving qualitative picture. The assumption of negligible forces of interaction (except during extremely short intervals of collision, “short” meaning short relative to the average interval between collisions) among the widely separated molecules, implies negligible storage of potential energy in the ideal gas. (Students should not lose sight of the fact that storage of potential energy in molecular interactions *does* play a signif-

ificant role in real gases and that large amounts of potential energy are associated with the latent heat of condensation to liquid state.) The transfer of heat to a gas at constant volume is then visualized as increasing the internal thermal energy of the gas by increasing the random kinetic energy of the molecules. If the molecules have no internal degrees of freedom, their kinetic energy is purely translational. If they can be excited into rotation, then rotational kinetic energy would be involved. If internal vibration is possible, then the molecules would possess both internal potential and kinetic energies of vibration that could be exchanged in collisions, etc. There is in addition, of course, the continual absorption and emission of photons in interaction with the sea of thermal radiation.

On the one hand these qualitative insights pave the way for understanding one of the principal failures of classical theory and initial successes of quantum theory (namely, the accounting for the observed anomalies in the specific heats of gases), and, on the other hand, to an appreciation of why one starts the most elementary kinetic theory with the consideration of point-mass particles, thus avoiding, at least at the beginning, the complexity imposed by molecular internal energy other than the purely translational. Without such qualitative background, the assumption of point-mass particles is unmotivated and carries no meaning to the students. They memorize it, if necessary, among the various inexplicable assumptions, but they regard it as still more unintelligible black magic.

- 10 A final valuable homework exercise, not strictly within the realm of kinetic theory of the ideal gas, nevertheless helps students register many essential ideas. This assignment might run as follows:
- (a) Describe in terms of motion and migration of molecules what it is that happens in evaporation of liquid water from a container that is open to the air. Supplement your verbal description with sketches, however crude the latter may be. Keep in mind the fact that some water molecules that have escaped from the liquid may return to it in their random motion and collisions. In terms of the kinetic model, explain how it is that all the liquid eventually evaporates as long as the container is not covered.
 - (b) Describe in similar terms what happens when water, with an air space above it, is put into a tightly closed container. In terms of the kinetic model, why is it that, after a relatively short time, no further net evaporation of water takes place under these circumstances? Do molecules cease leaving the water when the net evaporation ceases? The space above the liquid water is said, under these conditions, to be "saturated" with water vapor, and the contribution of water molecules to the total pressure of the gas phase on the walls of the

container remains constant. How does this contribution manage to remain constant despite the continual random motion of the molecules and the fact that some of them must return to the liquid phase during their wanderings?

- (c) Consider the situation in which a lump of sugar is placed in water. Although the process is slow in the absence of stirring, all the sugar eventually dissolves if the amount of water is fairly large. If the amount of water is relatively small, however, the dissolving ceases, some of the solid sugar remains, and the solution is said to be “saturated.” Describe in terms of the kinetic-molecular model, with the aid of sketches, what happens in these two cases, indicating how it is that equilibrium is achieved in one case and not in the other. Compare this situation with the evaporation discussed in parts (a) and (b).

11.3 HYDROSTATIC PRESSURE

Understanding the concept of “hydrostatic pressure” involves a good bit more than acquiring the definition “force per unit area.” A major step toward understanding resides in appreciation of the full significance of Pascal’s law: That the pressure at any point in a fluid is the same in all directions. The usual formal “proof” of this idea is given by examining the static equilibrium of a small volume of fluid, for example, one having a parallelepiped shape and a triangular cross section. This treatment is so abstract that, even if it is presented, very few students assimilate its physical implications. Many introductory courses now eschew such treatments entirely in order to save time for other subjects. This is a legitimate choice, but teachers must then remember that something very subtle and fundamental has been left out and must be prepared to help close the gap at some subsequent point.

The subtlety of the insight, and the fact that many individuals (including many active physicists) have not really acquired it, are indicated by the responses given to the following question: A container of the shape in Fig. 11.3.1 initially contains a uniformly dispersed mixture (or colloidal suspension) of two immiscible liquids of different densities (e.g., oil in water; cream in milk). As time goes by, the lower density fluid separates and collects in the throat of the container. How does the final pressure on the *bottom* of the container (after the separation is complete) compare with the initial pressure (when the fluids were uniformly dispersed)? Is it the same, greater, or smaller than the initial pressure?

Novices tend to say they do not know, but the majority of those who have had some physics but have never thought about such manifestations of fluid pressure tend to say that the pressure must remain the same. Those who suggest a line of reasoning rather than just making an intuitive guess say

that, since the weight of the fluid is unchanged, the total force on the bottom, and therefore the pressure on the bottom, must remain unchanged. This the reasoning given by many active physicists and physics teachers of the present day.

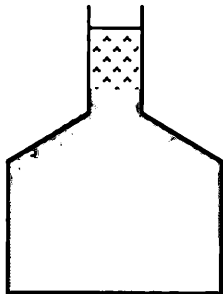


Figure 11.3.1 Initially uniform mixture of two immiscible fluids separates, and lower density fluid collects in throat of container. Does pressure on bottom of container change with separation of the fluids?

This reasoning is incorrect because the pressure on the bottom of a container of any shape other than purely cylindrical is *not* equal to the weight of fluid divided by the area of the bottom. Since the fluid pressure is the same in *all* directions at any point, the fluid at a wall exerts a force on the wall, and the wall, in turn, exerts a force on the fluid. The sloping walls of the container exert a downward component of force that influences the pressure on the bottom of the container. Since, on separation, the average density of the fluid in the central column is less than it was initially, the pressure at the bottom of the central column has decreased. The pressure at the level of any point along the sloping walls has also decreased; the downward force exerted by the sloping wall has decreased; the pressure has decreased all over the bottom.

Students can be helped to understand the physics of the phenomenon through the drawing of simple force diagrams, for example, force diagrams of a central column of fluid and of columns under the sloping walls. The drawing of such diagrams helps develop understanding of the physics and of the real meaning of fluid pressure, and the spiralling back to force diagrams helps strengthen the usually still uncertain grasp of the force concept (as discussed in Chapter 3).

An illustration of how an alteration in context can help deepen understanding: If, after discussion of the phenomena of fluid pressure, one turns students' attention to the pressure at the base of a vertically standing solid metal cone or cylinder, and asks what will happen to the pressure at the base of the cylinder if it expands under a uniform increase in temperature, many students (especially those who have memorized the formula $p = \rho gh$ without real understanding of where and when it applies) will say that the pressure at the base increases because the height increases. In the case of the solid, the pressure at the base is, of course, equal to the weight of the object divided by the area of the base. In the case of thermal expansion, the change in height has no effect, and the pressure at the base decreases (very slightly) because of the increase in area stemming from the lateral expansion.

Such contrasts in phenomenology play a powerful role in enhancing understanding, but few students can raise such questions on their own. They begin to do so only if they are led into the effort, gradually, and through many examples.

11.4 VISUALIZING THERMAL EXPANSION

Thermal expansion is another subject that is frequently skipped nowadays in introductory physics for the sake of other topics. This is just as legitimate as the skipping of fluid phenomena, but, again, teachers should be cognizant of certain widely prevalent gaps and misconceptions, the most important of which is the failure to visualize the overall effect of linear expansion of solids. (Note instance cited near the end of the preceding section.)

Many students fail to see that linear expansion in all directions means that all lengths, all areas, and all elements of volume in an object expand simultaneously. Thus they fail to see that the *circumference*, not just the diameter, of any circle drawn in the object increases in length. As a result, they fail to comprehend that a hole in a metal plate will increase in size on thermal expansion rather than decrease.

If thermal expansion is to be understood, students must be led to confront the qualitative phenomena in addition to the standard numerical exercises. In this instance, one must visualize the details of effects that cannot be seen directly; they transcend direct sense experience. Such exercises are essential in building up students' capacity for abstract logical reasoning and for using concepts as a basis for understanding more complex phenomena. Simple, everyday phenomena, such as thermal expansion, can provide very valuable contexts for building such capacity. It pays to invoke them, at least as minor digressions, without the expenditure of excessive time on numerical exercises such as those common in older texts.

11.5 ESTIMATING

A widely prevalent faculty complaint concerns students' unwillingness and resistance to doing simple calculations that involve estimating magnitudes of any kind. The thinking that goes into such calculations is, nevertheless, very sophisticated, however simple it may seem to experienced physicists. Students can be helped to develop the capacity, but only through being given the chance to practice in meaningful and interesting contexts. Since scaling and ratio reasoning are inextricably involved in almost all estimating, and, given the difficulty that the majority of students have with such reasoning (see Sections 1.6-1.12), teachers must recognize the underlying difficulty and provide the necessary guidance and simpler exercises where necessary. Only then will the capacity begin to develop.

A great deal of thought, effort, and imagination has gone into invention of fruitful problems of this variety—problems that can be readily incorporated into homework and tests in existing courses even though very few occur in standard texts. Teachers interested in drawing on such resources will find a rich harvest in various references in the bibliography, for example, Bartlett (1976-1979), Crane (1960, 1969a,b, 1970), Hobbie (1973), Kunz (1971), Lin (1982), Memory and Jenkins (1977), and Morrison (1963).

11.6 SIGNIFICANT FIGURES

Making students aware of the concept of “significant figures” and cultivating appropriate handling and reporting of numerical data was difficult enough before advent of the hand calculator. The difficulty has, of course, been significantly compounded by advent of the latter (play on words intended.) This does not reduce our responsibility for dealing with the idea effectively in elementary physics. An extensive treatise on the subject is not appropriate in this book, but I include a few brief recommendations based on personal experience.

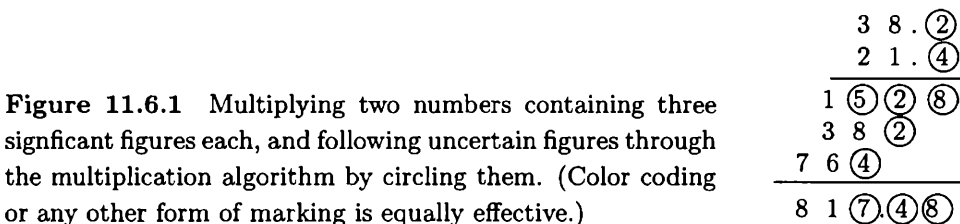
As in many other instances (e.g ratio reasoning), quick, remedial exercises at the beginning of course or laboratory work are ineffective. The concept must be brought out as direct experience is being accumulated. In making *all* measurements, however simple (even length measurements with a ruler), students should be required to record three values: Largest reasonable value, “best” value, and smallest reasonable value.

Out of such records, one then gradually elicits the basic scientific code, namely that the last figure in reported data is to be the first uncertain one. Students should be led to see that reporting figures beyond the first uncertain one is misleading and *untruthful*. (Most are impressed by the examination of how best to maintain truthfulness.) Such examination should include the different meanings of 3 in mathematics and 3.0, 3.00 in reporting results of measurements, as well as what is being made clear when one writes 3.00×10^5 rather than 300,000.

Having established the code for indicating where truthfulness ends in numerical data, one proceeds to examine the effect of numerical calculations involving addition, subtraction, multiplication and division. One advantage of the hand calculator is that it is not too laborious to calculate three values (largest, best, and smallest) from the three values available in the recorded data. The effect of subtraction is dramatically revealed. The effect in division must be carefully extracted because few students initially perceive that the largest result is to be obtained by dividing the largest by the smallest value rather than the largest by the largest. (This is part of the heritage from inadequate exploration of division and its consequences in the elementary grades.) One can deprecate the string of numbers emerging on the face of the

calculator by asking “What kind of an answer is 643.8567291 give or take 3 or 4?”

I would still not abandon the old written algorithms entirely unless we reach the point that students are totally incapable of resorting to them. The propagation of uncertain figures is still best illustrated by the old, well-known scheme of following them explicitly through the algorithm as illustrated in Fig. 11.6.1. The uncertain figures can be marked by circling or by color coding, and the useful rule of thumb (that the number of significant figures legitimate in the result of calculation is equal to the number of significant figures in the least precise number entering the calculation) can be brought out empirically.



The more rigorous development of the propagation of uncertainty through the following of “percentage error” is better left for spiralling back as greater confidence with ratio reasoning and percentages is cultivated; otherwise there is little residual impact other than desperate memorization.

11.7 PRECISION, ACCURACY, AND SIGNIFICANT DIFFERENCES

Recording three values (largest, best, and smallest) for each and every measurement the students make yields important strategic advantage besides the understanding of significant figures outlined in the preceding section. It leads naturally and intelligibly into the realm of statistical analysis without unintelligible assertions and calculations that have to be memorized or that are hidden within the black magic of the computer.

Students can be shown that we accept intuitively—as a primitive—the arithmetical average of the “best” values we have recorded as the *overall* best value. The reasonable range of uncertainty is given by the smallest and largest values emerging from the calculations. Two results differ significantly if their averages lie *outside* the range of uncertainty and do *not* differ significantly if the averages lie *within* the range of uncertainty.

At the same time, one can sensitize students to the difference between “precision” of the data on the one hand (low scatter of the measurements) and “accuracy” of the result (freedom from systematic error) on the other.

At the introductory level most students have been imbued with false notions about the “exactness” of science, and, through unfortunate laboratory experience (if any), have come to associate “error” with sin. (Here is another instance in which we have taken a widely used metaphor out of everyday speech and endowed it with greatly altered scientific meaning.) Encountering data and calculations in the simple, unhurried manner recommended above leads students to the perception that error, in the sense in which we use the term in science, is not sinful but is informative and revealing. They can come to appreciate Piet Hein’s little “Grook” :

*Knowing what
thou knowest not
is in a sense
omniscience.*

Given this primitive introduction to scatter and uncertainty, students can be led to appreciate the need for better and more sophisticated *quantitative* assessment of these aspects. Those who go on further can begin to understand the point of properties such as standard deviation and the role of probability even if the mathematical background is not developed rigorously, and they can begin to use the computer-based shortcuts intelligently instead of blindly. At highschool level and for those who do not go on beyond introductory college courses, I would argue that the beginning I have outlined is quite adequate and far superior intellectually to blind use of mystical numbers emerging from canned computer programs.

11.8 DISTRIBUTION FUNCTIONS

Many textbooks present diagrams or illustrations showing distribution functions, the most prevalent being the Maxwellian distribution of molecular velocities, the Gaussian error curve, and the spectral distribution of black body radiation. (The latter has recently returned to especial prominence because of its interesting connection to the 3°K residual cosmic radiation and the big bang cosmological model.)

Many teachers and textbook authors seem to be unaware that these diagrams are quite meaningless to students who have no conception of the meaning of a distribution function and who (even when they have some knowledge of histograms) do not distinguish it from a histogram.

The concept of the distribution function is subtle and difficult for most students because it has to be recognized as a *derivative*—a graph of the *slope* of the *cumulative* distribution function. If one is to deal with this idea honestly instead of speciously, it is necessary to bring students through the process of generating such functions and give them time to digest, absorb, and make mistakes.

I do not recommend pursuing the issue in introductory courses unless the teacher has selected a story line that requires the concept. I seek, however, to warn teachers immediately beyond introductory level that very few students have understood the meaning of a distribution function, and, if such functions are to be honestly invoked and utilized, one must take time to develop their meaning explicitly. Otherwise one is again cultivating blind memorization.

11.9 GUIDING INSIGHT AND INQUIRY IN INTRODUCTORY LABORATORY⁴

The usefulness and effectiveness of introductory laboratory have been bones of contention in physics teaching as far as one cares to go back in the literature. Laboratory instruction is costly, and, since its effectiveness has been difficult to substantiate compellingly,⁵ some responsible administrators have viewed it as a luxury we cannot afford. Yet most physicists have a deeply rooted, intuitive feeling that laboratory experience is essential to learning and understanding our subject, and they fight hard for maintaining such instruction in the face of frequently expressed doubts and occasionally formidable opposition. I have always found myself in the latter group even though I am compelled to recognize that much laboratory instruction is indeed ineffective and falls far short of the objectives to which lip service is invariably rendered. In this section, I seek to define a few major reasons for the admittedly frequent failure and to outline some modes of thinking that seem to promise greater effectiveness and firmer justification for maintaining the laboratory as an essential component of physics teaching.

The most common objectives explicitly voiced in connection with introductory physics laboratory (and also the ones most frequently implied by the instructions provided and the contexts chosen) are: (1) to verify or confirm laws, relations, or regularities asserted in text, class, or lecture; (2) to have some experience with actual physical phenomena; (3) to have the experience of, and develop some skill in, handling instruments and making significant measurements; (4) to have the experience of planning and doing experiments and thus encountering some of the “processes of science”; (5) to learn something about minimizing error and about the treatment and interpretation of experimental data. All of these are, to some reasonable extent, legitimate and desirable objectives. They are pervasive in existing curricula and have been so extensively and lucidly discussed⁶ that they require no further elaboration,

⁴This section is based on an article originally published in *The Physics Teacher* **31**, 278-282 (1993).

⁵See, for example, Kruglak (1952a), Kruglak (1952b), Brown (1958), Flansburg (1972), White (1979), Moreira (1980), Toothacker (1983), Long, et al (1986).

⁶See, for example, Michels (1962), King (1966), Ivany and Parlett (1968), Prescott and Anger (1970), Shonle (1970), Reid and Arsenau (1971), Swartz and Zipfel (1972), Reif and St. John (1979), Robinson (1979), Phillips (1981), Kuehl, et al (1984).

I shall accept them as a given and shall not lengthen the present discussion by dwelling on their implementation.

It is quite apparent, however, that, in widespread implementation of laboratory instruction ostensibly based on these objectives, the students still emerge with levels of insight and achievement deprecatingly described in the literature cited earlier and with little more than what Whitehead (1929) describes as “inert ideas”, i.e., “ideas that are merely received into the mind without being utilised, or tested, or thrown into fresh combinations.” What can be done to help students to deeper learning and insight through utilizing, testing, and throwing into fresh combinations the ideas they encounter in the laboratory? I wish to suggest that we might make at least some progress in this direction by guiding students into utilizing one or more of the following criss-crossing and overlapping modes of inquiry and critical thinking:

- 1 Observing phenomena qualitatively and interpreting observations.
- 2 Forming concepts as a result of observations
- 3 Building and testing abstract models in the light of observation and concept formation.
- 4 By subjecting a piece of equipment to close examination in context, figuring out how it works and how it might be used (rather than simply being told how it works and what it is supposed to do.)
- 5 Deciding what to do with a piece of equipment as well as deciding how many measurements to make and how to handle and present the data.
- 6 Asking or pursuing “How do we know. . . ?, Why do we believe. . . ? What is the evidence for. . . ?” questions inherently associated with a given experiment.
- 7 Explicitly discriminating between observation and inference in interpreting the results of experiments and observations.
- 8 Doing hypothetico-deductive reasoning (i.e., asking and addressing “What will happen if. . . ?” questions) in connection with the laboratory situations. This includes visualizing, in the abstract, the effect of changing relevant variables or boundary conditions, visualizing outcomes in extreme or limiting conditions, and, where possible, forming an a priori hypothesis and then testing it by performing an appropriately designed experiment.

I hasten to make no claim that this list is unique, complete, or exhaustive. I mean it to be illustrative of some important possibilities, and I invite teachers to refine and amplify it with their own favorite modes and to their hearts' content. It should be noted, however, that only two or three of these

modes of reasoning are explicitly articulated in the statements of objectives in the literature that was cited earlier, and most instructions and manuals supplied to the students seem to assume that the sought for insights will develop automatically from the prescribed experience without additional guidance.

It has long been clear that tightly structured and directed laboratory experiments are dull and demoralizing for the students and generate little in the way of concept development or physical understanding. It is also clear that the other extreme of completely unstructured situations, in which students are supposed to conduct their own observations, inquiry, and final syntheses, are also ineffective [c.f., Spears and Zollman (1977)]. Very few students in introductory courses have had the opportunity to develop the very sophisticated stance that goes with such self sustained and systematic inquiry. (This is quite analogous to situations in which students are let loose, without guidance, on elaborate computer programs or dialogs and are expected to pursue fruitful exploration on their own. While experienced physicists may proceed to explore such programs with pleasure and excitement, most students simply sit and stare without having any idea of how to start, and, even if they get started in some fashion, fail to make significant progress or achieve significant insight without help or guidance.)

The problem is to provide students with enough Socratic guidance to lead them into the thinking and the forming of insights but not so much as to give everything away and thus destroy the attendant intellectual experience. (I deliberately use the word "guidance" to imply a distinction between this mode and the more conventional modes of providing instructions and answers.) Providing suitable Socratic guidance, however, is more of an art than a science, and it is critically dependent upon a number of factors unique to the situation in which it is to be employed: The context that has been set by the teacher; the vocabulary that has been used by the textbook and the teacher; the way in which concepts have been developed and presented; the nature of the test questions to which students are being exposed; the level of sophistication of the group of students being addressed. What I seek to do in the following is to give a few specific examples of the kinds of thinking that can be fruitfully cultivated in simple experiences and with commonly occurring apparatus. How this thinking is to be elicited Socratically without giving too much away is being left up to the individual teacher.

11.10 EXAMPLES OF CULTIVATING INSIGHT AND INQUIRY IN THE LABORATORY

The difficulties experienced by students in mastering the law of inertia and the concept of "force" and the robust preconceptions with which they approach mechanical phenomena have been extensively discussed in the literature and are widely appreciated by teachers. Qualitative hands-on experience in the

laboratory furnishes an effective way of helping many students overcome these difficulties. Examples of the questions students can be led to address through such experience are given in Sections 3.10 through 3.12 of this book, and a Socratically oriented laboratory aimed at the same objectives is described in considerable detail by Hake (1992). Such sequences illustrate application of modes of inquiry 1 and 2 in the list given above. Hypothetico-deductive reasoning can now be invoked by adding “What will happen if . . . ?” questions (item 8) concerning situations which the students must imagine and deal with in the abstract. Finally, students should be asked to distinguish explicitly between the observations that were made and the inferences drawn, illustrating application of mode 7.

Following the pattern outlined in detail in Arons (1977), students can be led to accumulate and pool observations, over whatever time period may be necessary, of the illumination exhibited by the moon. At the same time they keep track of the angular separation between the sun and the moon. (An indoor laboratory is, of course, completely unnecessary.) In the process they address the question “Why do we believe the moon shines by reflected sunlight?” and also invent the model by means of which they can keep track of the entire sequence of events. (Only as the model is being formed from the observations does the visual aid of the ball and beam of light become an appropriate adjunct. If it is introduced a priori, by assertion from authority, it simply destroys the learning experience.) The “Why do we believe. . . ?” question should be posed at the start and should be the principal goal of the investigation. After the model has been generated, it can be tested by predictions concerning configurations that may not have been observed directly. Addressing questions such as “Would you expect to see a full moon rising at midnight? Why or why not?” cultivates hypothetico-deductive reasoning. Finally, students should be led to distinguish between the observations that were made and the inferences drawn from the observations. This episode illustrates applications of modes of inquiry 1, 3, 7, and 8.

A similar pool of outdoor observations can be put together by a class observing, over a period of weeks or even months, the location along the horizon of rising of the sun, location of setting of the sun, angle at which the sun crosses the horizon in rising or setting, variation of the length and orientation of the shadow of a vertical stick over the course of a day and as days go by. In the process they can address questions such as the relevance of the observations to the meaning of the terms “local noon,” “local midnight,” “local celestial meridian,” “north-south direction.” They can address the question as to whether or not the sun ever passes “directly overhead” at their location. As observations accumulate, students can transfer the picture they are forming to a laboratory celestial sphere apparatus. (The latter should be introduced only as it becomes meaningful through the observations and not as an a priori assertion.) At this point students should be led to see that, in the light of the observations available, it is impossible to discriminate between a geocentric

and heliocentric model. This experience readily invokes application of modes of inquiry 1, 2, 3, 4, 7, and 8 if appropriate questions are raised at appropriate points. Furthermore, it lays the basis for a serious examination of the grounds on which we subsequently accept the heliocentric model.

Sequences of hands-on laboratory experience for developing the concepts of “electric circuit,” “electric current,” “conductor,” “nonconductor,” “resistance,” as well as the model for what happens in simple resistive, direct current circuits, are outlined by Arons [(1977), Chapter 9], McDermott and Shaffer (1992), and McDermott, et al (1996). If conducted Socratically with opportunity for group discussion and synthesis, and with subsequent questions and problems that utilize the model for predicting what will happen in new configurations, the situation is very rich in opportunities to invoke modes of inquiry 1, 2, 3, 7, and 8. This is also an excellent opportunity for encouraging students to invent configurations of their own and to ask their own “What will happen if . . . ?” questions. They can challenge each other with such questions in group discussions.

An important “How do we know . . . ?” conceptual experience can be conveyed by leading the students to construct a simple experiment aimed at answering the question “How do we know that the quantity $I\Delta V$ (the power supplied to a resistive circuit) is dissipated in a pure resistor and is not converted to other forms of energy (as it is if the power is supplied to a motor.)” See Sects. 7.6 and 7.11 in this book for a discussion of this question. Such an experiment can invoke modes of inquiry 4 and 5 if the students are given simple equipment for which they must figure out the point and function. Careful Socratic guidance is essential. This investigative structure can significantly enrich otherwise very routine, and essentially sterile, Ohm’s law experiments.

Many laboratories possess ballistic pendulums and a rotator apparatus for measuring centripetal force. Such traditional equipment is well adapted for invoking modes 4 and 5 as well as for leading students to decide what measurements to make and how to acquire and handle the data. Instead of being told what measurements to make, how many, and how to handle the data, they should be required to make their own decisions, allowed to do these things badly on a first go-around, and encouraged to arrive at improvements through discussion and critique of what has been done. They should not be given low grades or “punished” for doing these things badly to begin with. We all learn far more from our own mistakes than we do from someone else’s correct instructions.

Many laboratory courses include an experiment in measuring the “mechanical equivalent of heat” because the experiment is relatively easy to perform and maintain and because it invokes many of the conventional laboratory objectives that I have earlier accepted as given. The literature is replete with descriptions of such experiments, and improvements are continually being described [e.g. Galloway and Wilson (1992) and Weber (1992).] Such papers have, in recent years, brought forth letters of criticism, such as that by Hooper

(1993), in which it is contended that “The unnecessary introduction of the old-fashioned calorie causes the problem that the described experiments then seek to solve; namely, ‘How many joules are there in a calorie?’ Contemporary handling of this experiment requires the recognition that both heat and mechanical energy are measured in joules, thus eliminating the problem.” Hooper goes on to suggest that the experiment be converted into one of measuring the specific heat of water.

The way in which such experiments are usually presented to students and the criticism, as expressed by Hooper, both conspire to render the experiment completely “inert” in Whitehead’s sense. They reveal a profound misunderstanding of what the experiment embodies. There is indeed little point in finding out, for its own sake, how many joules there are in a calorie. Concealed from students, however, in most such cases is the profoundly important conceptual question “How do we *know* that heat can be regarded as a form of energy and that it can therefore be measured in joules?” This question is rarely, if ever, brought forth in student minds.

Feynman (1963), as is usually the case, gives a beautiful statement of what is involved: “There is a fact, or if you wish, a law, governing all natural phenomena that are known to date. There is no known exception to this law—it is exact so far as we know. The law is called the conservation of energy. It states that there is a certain quantity, which we call energy, that does not change in the manifold changes which nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says there is a numerical quantity which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a strange fact that we can calculate some number and when we finish watching nature go through her tricks and calculate the number again, it is the same.”

The basic problem, of course, is to discover how such numbers must be calculated, and that is not obvious to the students any more than it was obvious to the scientists of the first half of the 19th century. The question to which students should be guided is “How do we know that heat can be regarded as a form of energy?” The whole idea is rendered inert if the correspondence is asserted *ex cathedra* as an end result not worthy of inquiry. The real point of the mechanical equivalent of heat experiments is not to find how many joules there are in a calorie but to show that work dissipated (measured in joules) is directly proportional to the simultaneously produced thermal effect as measured independently in some other unit such as the calorie or B.T.U. It is because of this direct proportion (originally established by Joule and Mayer) that we come to recognize heat as a form of energy and can calculate numbers such those Feynman describes. Leading students to this “How do we know . . . ?” question and helping them see its conceptual significance sharpens their ability to start asking their own “How do we know . . . ?” questions and keeps the laboratory experience from simply being an inert end result. There is rich opportunity here for invoking modes 6 and 7.

Goldberg and McDermott (1987) (see also Sect. 9.18 in this book), in investigating the conceptions with which students emerged from a laboratory experiment in which they confirmed the Gaussian lens equation for image formation by a thin converging lens, showed that the many of the students had acquired no understanding of the function of the screen, no understanding that each point on an object radiates light in all directions, no understanding that every point on the lens contributes to every point on the image. The most important physical ideas had not been exposed, and the one investigated was rendered completely inert. Virtually nothing had happened along the lines of modes 1, 2, 5, 6, 7, or 8, although, with appropriate guidance, the situation is rich with opportunity for exercising all of these modes.

I have tried to illustrate, in a few specific cases, how, with judicious guidance, with awareness of the modes of inquiry that can be invoked, and with thorough understanding of the physics involved, one can endow even conventional laboratory experiments with a far greater degree of intellectual richness than they possess in their average levels of implementation. To determine whether students exposed to such laboratory experience achieve higher levels of insight and understanding than students without such experience, one would have to use tests that probe for the concepts and modes of reasoning being cultivated. Such probing cannot be achieved by testing with conventional end-of-chapter numerical problems; such testing will, as always, confirm the null hypothesis.

11.11 EXAMPLES OF MATHEMATICAL PHYSICS FOR GIFTED STUDENTS

A profoundly important aspect of intellectual development in students who are likely to become physicists (or engineers who use physics at a sophisticated level) resides in being able to extract subtle physical interpretation from analytical results. This is, again, a matter of practice and experience, but very little such experience is generated in standard problems in the textbooks until later, more advanced, levels of study. It is highly desirable to expose gifted students to such practice at the earliest possible moment, but, at introductory physics level, it is necessary to invoke situations that are sufficiently simple to be analyzed with the conceptual and mathematical tools then available.

Such situations must be sufficiently rich phenomenologically to offer non-trivial problems of interpretation, that is, the situation must not be intuitively obvious, and the mathematical solution must be essential for penetration of the physics. Such problems are not easy to find; a few that I have found valuable are illustrated in the following. It must be emphasized that these problems are not original and are quite well known. They do not, however, generally occur in textbooks, and the key to their use is not so much in the problems themselves as in how they are presented to the student. The basic approach in each

instance is to have the student set up the problem, obtain the formal solution, and then interpret the solution by extracting its physical implications. Playing with actual apparatus (where this is feasible) can follow the experience of doing the abstract mathematical physics. Reversing the procedure is perfectly possible, but it destroys the mathematical-physics experience.

Problem A A bob of mass m is attached directly to a spring of finite length having spring constant k , and the system is whirled around in a horizontal circle at angular velocity ω as shown in Fig. 11.11.1. The axis of rotation is the vertical line through the end of the spring.

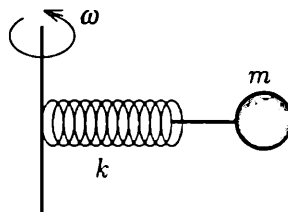


Figure 11.11.1

At the initial, relaxed condition the distance of the center of mass of the bob from the axis of rotation is denoted by r_o , and, at angular velocity ω , this distance (the radius of the circular motion) becomes r . Examine the dynamics of this system from the point of view of determining the way in which the equilibrium value of r varies with imposed values of ω . Interpret the mathematical result by describing the physical effects predicted. Do not try to treat this situation as one in which r varies with time; just find how values of r depend on steady values of ω .

Solution: Since, at any final radius r the spring has been stretched by an amount $r - r_o$, the centripetal force exerted on the bob is given by $k(r - r_o)$. The dynamical equation is therefore

$$k(r - r_o) = mr\omega^2 \quad (11.11.1)$$

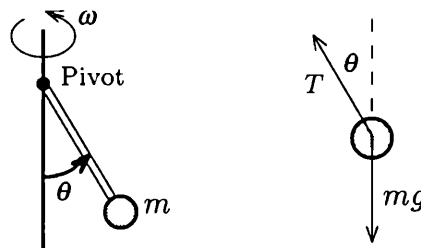
Solving for r , we obtain

$$r = \frac{r_o}{1 - \frac{m\omega^2}{k}} \quad (11.11.2)$$

Interpretation: Equation 11.6.2 shows that, starting from zero angular velocity, r increases as ω increases as would be expected. What is unexpected, however, is that the system blows up as ω approaches the angular frequency of the natural oscillation of the mass on the spring.

Problem B A pendulum bob of mass m is fixed at the end of a thin rigid rod of length L . The mass of the rod is negligible compared to that of the bob. The bob and rod form a conical pendulum that can be rotated at angular velocity ω around a vertical axis through the pivot at the end of the rod (Fig. 11.6.2). Obtain the analytical expression for the angle θ assumed by the rod at various steady angular velocities ω and interpret what it predicts about the physical behavior of the system.

Figure 11.11.2 Conical pendulum (bob of mass m on rigid rod of negligible mass) is rotated at angular velocity ω around a vertical axis through a pivot at the end of the rod. At steady angular velocity, the rod takes up an angle θ from the vertical.



Solution: From the force diagram in Fig. 11.11.2, we have

$$T \cos \theta - mg = 0 \quad (11.11.3)$$

and

$$T \sin \theta = mL\omega^2 \sin \theta \quad (11.11.4)$$

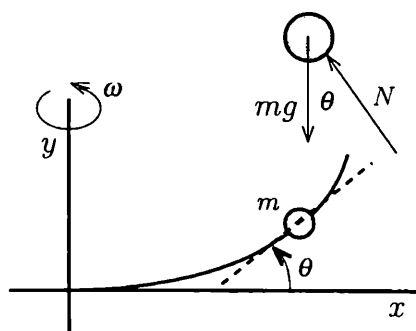
Combining Eqs. 11.11.3 and 11.11.4, gives

$$\cos \theta = \frac{g}{L\omega^2} \quad (11.11.5)$$

Interpretation: Equation 11.11.5 shows that, for $\omega < \sqrt{g/L}$, $\cos \theta$ would be greater than 1. Thus θ would be imaginary and consequently meaningless. The solution implies that the pendulum does not begin to swing out from its zero position until the angular velocity ω exceeds the natural angular frequency of the pendulum. This is indeed the case, as can be readily confirmed by an experimental observation.

Problem C

Figure 11.11.3 A rigid wire is bent into the form of the power function $y = ax^n$, where n takes on integer values beginning with unity. The wire is rotated around the vertical y -axis at angular velocity ω . A frictionless bead of mass m can slide along the wire and take up equilibrium positions at various steady values of ω .



Obtain the analytical expression for the equilibrium x -coordinate of the bead as a function of steady angular velocity ω and interpret the results physically. Note that the results are quite different for different values of n ; pay especially careful attention to interpretation of results for the cases where $n = 1$ and $n = 2$. [Note that, since both x and y are to have the physical dimension of length, the constant a must have dimensions of $(\text{length})^{-n+1}$].

Solution: From the force diagram in Fig. 11.11.3, we have

$$N \cos \theta - mg = 0 \quad (11.11.6)$$

and

$$N \sin \theta = m x \omega^2 \quad (11.11.7)$$

Combining Eqs. 11.11.6 and 11.11.7 gives

$$\tan \theta = \frac{x \omega^2}{g} \quad (11.11.8)$$

Since, from the shape of the wire, we have

$$\tan \theta = n x^{n-1} \quad (11.11.9)$$

we finally obtain

$$x^{n-2} = \frac{\omega^2}{nag} \quad (11.11.10)$$

Interpretation: For values of n greater than 2, the equilibrium value of x is stable and increases, as would be expected, with increasing ω . For the case of the straight wire ($n = 1$), the equilibrium position is given by $x = ag/\omega^2$, but this is an *unstable* equilibrium in which the bead would slide inward on one side and outward on the other.

The most interesting case is that for the parabolic shape $n = 2$. Here, when the angular velocity is equal to $\sqrt{2ag}$, the bead will stay wherever it is placed along the wire. This corresponds exactly to the situation in the rotating basin of water in which, at a given steady angular velocity, the water surface adjusts itself to the appropriate parabolic shape, and every particle of water then occupies an equilibrium position.

11.12 CHAOS

One of the deepest scientific insights attained in recent years, owing largely to the possibilities opened by the computer, has been the realization that classical mechanics is intrinsically indeterminate because of ever-present non-linearity. This indeterminacy is fundamentally different from that of quantum mechanics, but it is universal and just as deeply significant. What this insight reveals is that Laplace was wrong: Given the values of position and velocity of every particle in the universe, it is not in fact possible (even in a classical Newtonian universe) to predict the subsequent history of that universe. Regardless of how closely together initial conditions are taken, the solutions diverge exponentially when the governing equations are nonlinear.

This insight is one to which we should try to expose our students, especially those in engineering-physics courses, and the ubiquity of adequate computers helps make this possible. The mapping of solutions in phase space also enhances grasp of the meaning of phase space and prepares the way for subsequent, more sophisticated levels of study.

How far one then goes with examination of still deeper aspects (e.g., strange attractors, “dimensionality,” connection to fractals, etc.) is a matter of available time and of judgement on the part of the teacher. I believe that we should at least aim for the first level of insight concerning the divergence of solutions regardless of the “exactness” of initial conditions.

Chapter 12

Achieving Wider Scientific Literacy

12.1 INTRODUCTION¹

That public understanding of science, or scientific literacy in general, is in a lamentable state is an old story. Magazine articles, educational literature, and pronouncements of scientific associations all through this century and well back into the last bear testimony to this. Scientists of fifty years ago, returning to academic work after time spent in research or serving in World War II, were especially determined to help college students acquire a better sense of the nature, power, and limitations of scientific thought, as well as a better understanding of the interactions between science and society. Since then, the escalating impact of science and technology on moral, ethical, political, and societal problems has only continued to enhance the urgency of the problem of education and to heighten the pertinence of the liberal education objective.

For years, meetings of scientific societies reverberated, and pages of educational journals were filled, with descriptions of new courses that had been designed to lead nonscience majors to greater scientific literacy. Almost every report presented or published was accompanied by the results of “evaluations” of student answers to tendentious questionnaires, the answers invariably demonstrating how much the students loved the course, valued the learning experience, and appreciated the instructor’s enthusiastic efforts on their behalf. With but a *very* few exceptions, however, these courses vanished and were rapidly succeeded by “more up-to-date” but essentially identical—and equally evanescent—versions, also accompanied by enthusiastic student testimonials.²

These numerous attempts have had very little impact on scientific literacy. Those who were students in such courses, and responded so favorably on the

¹This chapter is based on a paper originally published in *Daedalus*, Spring 1983.

²I wish to make it clear that I am not deprecating student opinion as such. I do question, however, the specious use of such opinion by faculty who have failed to provide the students with an adequate frame of reference from which to judge what has and has not been learned.

questionnaires, show little or no understanding of science and of its interactions with society. In retrospect, most say that they enjoyed their course very much, but recall nothing of what they were supposed to have learned. (It has been sardonically suggested that, this being the almost universal outcome, perhaps we should direct our efforts to devising courses still more enjoyable and still easier to forget.) Yet the clamor for making science a more effective component of liberal education continues unabated, and with an urgency indicative of how little past efforts have achieved.

The notion that understanding of science can be achieved by purely verbal inculcation seems to me to be a principal source of failure. Experience makes it increasingly clear that exclusively verbal presentations—lecturing to large groups of intellectually passive students and having them read text material—leave virtually nothing in the students' minds that is permanent or significant. Much less do such presentations help the student attain what can be considered the marks of a scientifically literate person. Since such marks, however, underlie the contentions and recommendations in this chapter, it is well to stop at this point in order to enumerate some of them.

12.2 MARKS OF SCIENTIFIC LITERACY

I suggest that an individual who has acquired some degree of scientific literacy will possess the ability to:

- 1 Recognize that scientific concepts (e.g., velocity, acceleration, force, energy, electrical charge, gravitational and inertial mass) are invented (or created) by acts of human imagination and intelligence and are not tangible objects or substances accidentally discovered, like a fossil, or a new plant or mineral.
- 2 Recognize that to be understood and correctly used, such terms require careful operational definition, rooted in shared experience and in simpler words previously defined; to comprehend, in other words, that a scientific concept involves an idea first and a name afterwards, and that understanding does not reside in the technical terms themselves.
- 3 Comprehend the distinction between observation and inference and discriminate between the two processes in any context under consideration.
- 4 Distinguish between the occasional role of accidental discovery in scientific investigation and the deliberate strategy of forming and testing hypotheses.
- 5 Understand the meaning of the word "theory" in the scientific domain, and have some sense, through specific examples, of how theories are formed, tested, validated, and accorded provisional acceptance; recognize, in consequence, that the term does not refer to any and every

personal opinion, unsubstantiated notion, or received article of faith and thus, for example, to see through the creationist locution that describes evolution as “merely a theory.”

- 6 Discriminate, on the one hand, between acceptance of asserted and unverified end results, models, or conclusions, and, on the other, understand their basis and origin; that is, to recognize when questions such as “How do we know. . . ? Why do we believe . . . ? What is the evidence for. . . ?” have been addressed, answered, and understood, and when something is being taken on faith.
- 7 Understand, again through specific examples, the sense in which scientific concepts and theories are mutable and provisional rather than final and unalterable, and to perceive the way in which such structures are continually refined and sharpened by processes of successive approximation.
- 8 Comprehend the limitations inherent in scientific inquiry and be aware of the kinds of questions that are neither asked nor answered; be aware of the endless regression of unanswered questions that resides behind the answered ones.
- 9 Develop enough basic knowledge in some area (or areas) of interest to allow intelligent reading and subsequent learning without formal instruction.
- 10 Be aware of at least a few specific instances in which scientific knowledge has had direct impact on intellectual history and on one’s own view of the nature of the universe and of the human condition within it.
- 11 Be aware of at least a few specific instances of interaction between science and society on moral, ethical, and sociological planes.
- 12 Be aware of very close analogies between certain modes of thought in natural science and in other disciplines such as history, economics, sociology, and political science; for example, forming concepts, testing hypotheses, discriminating between observation and inference (i.e., between information from a primary source and the interpretations placed on this information), constructing models, and doing hypothetico-deductive reasoning.

I hasten to indicate that this list is neither exhaustive nor prescriptive. It illustrates some of the insights that I believe characterize scientific literacy and that I find most college undergraduates, given time and opportunity, and having the willingness to exert some intellectual effort, can encompass. Readers will have valid modifications, preferences, and priorities of their own. These can be interpolated and examined in light of the following discussion,

which I will confine to efforts being made in schools, colleges, and universities to upgrade scientific literacy.

12.3 OPERATIVE KNOWLEDGE

Researchers in cognitive development describe two principal classes of knowledge: Figurative (or declarative) and operative (or procedural) [Anderson (1980); Lawson (1982)]. Declarative knowledge consists of knowing “facts”; for example, that the moon shines by reflected sunlight, that the earth and planets revolve around the sun, that matter is composed of discrete atoms and molecules, that animals breathe in oxygen and expel carbon dioxide. Operative knowledge, on the other hand, involves understanding the source of such declarative knowledge (How do we know the moon shines by reflected sunlight? Why do we believe the earth and planets revolve around the sun when appearances suggest that everything revolves around the earth? What is the evidence that the structure of matter is discrete rather than continuous? What do we mean by the names “oxygen” and “carbon dioxide” and how do we recognize these as different substances?) and the capacity to use, apply, transform, or recognize the relevance of the declarative knowledge in new or unfamiliar situations.

To develop the genuine understanding of concepts and theories that underlies operative knowledge, the college student, no less than the elementary school child, must engage in deductive and inductive mental activity coupled with interpretation of personal observation and experience. Unfortunately, such activity is rarely induced in passive listeners, but it can be nurtured, developed, and enhanced in the majority of students providing it is experientially rooted and not too rapidly paced, and providing the mind of the learner is actively engaged.

There is increasing evidence that our secondary schools and colleges are not doing very well at cultivating operative knowledge in any of the formal disciplines, and that the teaching of science is not unique in this respect—although the failures in science are more immediately obvious [Arons (1976); Chiappetta (1976); McKinnon and Renner (1971)]. Consider some specific illustrations:

- 1 Almost any individual (child, student, or adult) who is asked about the origin of the light coming to us from the moon will respond with the assertion that the moon shines by reflected sunlight. When one asks, however, for the evidence for this conclusion, one very rarely obtains a meaningful or logical response. The knowledge is purely declarative and has been received from authority without accompanying evidence or support. It is interesting to note a deeply related misconception: Most people, including nonscience college faculty, if asked how they account for the unilluminated portion when they see a bright crescent moon,

respond that the dark portion is the shadow of the earth. Very few people have ever watched the moon in its changing phases and taken the intellectual step of noting the simultaneous location of the sun. It is perfectly possible to lead young children to full understanding of what is going on and why we conclude that "the moonlight is the sunlight" (see Tennyson's "Locksley Hall Sixty Years After"), but this is very rarely done.

- 2 In a more subtle and sophisticated context, virtually any individual will tell you that the earth and planets of our solar system revolve around the sun. Most people do not even see anything paradoxical about this because, unlike the ancients, few of us now have occasion to sleep out under the sky and watch the procession of the celestial bodies. If asked for the evidence, for the reasons why we accept a helio- rather than a geocentric model, the vast majority, including college science majors, react only with dismay or embarrassment. A few might mutter something memorized and unintelligible about "stellar parallax," but even these have no realization that the Newtonian picture was firmly accepted long before stellar parallax was actually observed, that the observation was simply a confidently expected confirmation, and that a proponent of geocentrism could readily invent some refinements that would incorporate stellar parallax in the geocentric model. Thus most individuals have "learned" what Whitehead (1929) describes as an inert "end result." They possess only declarative knowledge received from authority, and they have no understanding of the first grand synthesis provided by modern science. They are probably even less sophisticated than their medieval counterparts, who would have put forth the geocentric model but would have qualified it as only "saving the appearances" rather than representing a final Truth. Such modern-day reactions of purely declarative knowledge are typical with respect to many other aspects of science, and, I submit, are not what we have in mind when we speak of an "understanding of science."
- 3 An example drawn from experience with both pre- and in-service elementary school teachers in undergraduate science courses (the two groups turn out to be indistinguishable in their levels of understanding of science subject matter): Somewhere in their general science courses in the schools, or in other circumstances, they had all heard expositions about "electrical circuits," had seen diagrams in books or on chalkboards, and listened to assertions of the facts and concepts of current electricity. When they are given a dry cell, a length of wire, and a flashlight bulb and are asked to get the bulb lighted, they almost invariably do one of the following things: they either hold one end of the wire to one terminal of the battery and touch the bottom of the bulb to the other end of the wire, or they connect the wire across the terminals (i.e., short

the battery) and hold the bulb on one battery terminal. They have no sense of the two-endedness of either the battery or the bulb; few notice that the wire gets hot when connected across the battery terminals, and fewer still infer anything from the latter effect. It takes 20 to 30 minutes before someone in the class discovers, by trial and error, a configuration that lights the bulb. (Seven-year-old children, when confronted with the same situation, go through exactly the same initial steps, and 20 to 30 minutes elapse before someone gets the bulb lighted.) Lacking the synthesis of actual experience into the concept of “electrical circuit,” the college students, despite the words they “know” and the assertions and descriptions they have received as passive listeners, have no more understanding of the ideas involved than the seven-year-old approaching the phenomenon *de novo*. Purely verbal inculcation has left no trace of genuine knowledge or understanding. Such is the outcome of the majority of our present modes of science instruction.

12.4 GENERAL EDUCATION SCIENCE COURSES

The majority of college courses that purport to cultivate scientific literacy in the nonscience major tend to fall into two principal classes: Courses that in one quarter, one semester, or even one year attempt to give students an insight into the major achievements of a science (e.g., in physics, everything from Galileo and Newton through the laws of thermodynamics, relativity, quantum mechanics, and current particle physics); and courses that focus on some narrower topical area such as the energy crisis, spoliation of the environment, the application of science to military problems, ethical and moral questions lying behind modern advances in molecular biology, philosophical questions posed by relativity and quantum mechanics, and so on.

Courses in the first category have been invented and reinvented in essentially the same form countless times ever since general education curricula sought to provide courses addressed to nonscience majors. Despite pretensions to being substantive and not merely “surveys” and despite the always glowing student “evaluations,” these courses have had so short a half-life and so little effect on the generations of students subjected to them, that they are still being reinvented to fill the persisting vacuum. Young scientists, completely unaware of past experience, seem to think the vacuum is there because this mode has never been tried before and that the solution lies in presenting the material in their own specially enthusiastic and impeccably lucid way. The truth is that the vacuum is there because this mode has no prospect whatsoever of educational success, yet its proponents continue to justify it on the ground that students are given a “feeling” for the content of science and the nature of modern scientific thought and that they now “know” something about current scientific progress. Meanwhile, complaints about the lack of scientific literacy continue to escalate.

Such efforts founder—as their replications will continue to founder—first, because they invariably subject students to an incomprehensible stream of technical jargon that is not rooted in experience accessible to the learner; second, because the subject matter is poured forth much too rapidly and in far too great a volume for significant understanding of ideas, concepts, or theories to be generated and assimilated. The pace makes difficult, if not impossible, the development of a sense of how concepts and theories originate, how they come to be validated and accepted, and how they connect with experience and reveal relations among seemingly disparate phenomena. Both the pace and the volume preclude any meaningful reflection on the scope and limitations of scientific knowledge or of its impact on our intellectual heritage and view of man's place in the universe. The “stream of words” courses have not solved, and will not solve, our educational problem, however handsomely illustrated the texts and however liberally salted they may be with allusions to pollution, ethics, energy crises, black holes, or Kafka.

Courses in the second category, although supposedly narrower in scope, nevertheless suffer from related difficulties. It seems to me that intellectual integrity would demand that students acquire some genuine comprehension of the scientific concepts, theories, and insights underlying the great topical problem being examined, and that students should not be encouraged to discourse vacuously on matters they essentially do not understand. With students who already have the requisite conceptual background, one can, of course, enter these discussions directly. But with students who have no notion of what “energy” means (many regard it as some kind of material substance) and no comprehension, however qualitative, of the restrictions imposed on us by the laws of thermodynamics; with students who have no basis for belief in the discreteness of the structure of matter (knowing only a string of names such as “atom,” “molecule,” “nucleus,” “electron” that have been thrown at them by assertion without examination of any of the experiential evidence and reasoning underlying the names); with students who have no idea what is meant (and what is not meant) by “electrical charge”; with students who have no understanding of the grounds on which we accept the proposition that the earth and planets revolve around the sun; and with students who are still Aristotelian in their use of teleological locutions and in their unawareness of the law of inertia—with such students it is intellectually specious and dishonest to pursue the initial discussion without first helping them form and understand the essential prior concepts. Indeed, once they see where intellectual integrity lies and what they must understand to talk meaningfully and intelligently about the original problem, few students object to the digression necessary for understanding the underpinnings.

Such backtracking to the necessary scientific understanding, however, drastically reduces the amount of coverage. To the best of my knowledge, very few courses have made the sacrifice, and the students emerge with no more understanding of the scientific concepts or of the nature and limitations of

scientific thought than do the victims of courses in category one.

What is the alternative? It seems to me that it is essential to back off, to slow down, to cover less, and to give students a chance to follow and absorb the development of a small number of scientific ideas at a volume and pace that make their knowledge operative rather than declarative. Depending on the time available, they might, for example, be led through one or more of the following questions:

- 1 Why do we believe the earth and planets revolve around the sun? In what context of concept and theory is this picture “correct”?
- 2 Why do we believe that matter is discrete rather than continuous in structure? What is the *evidence* behind belief in atoms and molecules?
- 3 What do we mean by the term “electrical charge”? How does the concept originate? Is “charge” some kind of substance? What is meant by “like” charges; in what sense is the word “like” being used? On what grounds do we believe that there are no more than two varieties of electrical charge? What (hypothetical) evidence would force us to conclude we had discovered a third variety? What is the evidence that electrical charge plays a fundamental role in the structure of matter?
- 4 Why do we believe that atoms have discrete structure on a subatomic scale? What is some of the evidence? (Ex cathedra assertion of terms such as “electron” and “nucleus” is not evidence at all and cultivates only declarative knowledge; yet much teaching of “science” is done in this fashion.) What experiments and observations lead to (and reinforce) creation of the concept “electron”? What is the evidence that such an entity is a universal constituent of matter? What is the evidence that it is subatomic? What role does the electron play in atomic structure?
- 5 Since we can take thermal energy out of the atmosphere or the ocean without violating the conservation principle, why is it that we cannot solve energy shortages by using the atmosphere or ocean as energy sources?
- 6 In what way does Einstein’s special theory of relativity alter our fundamental conceptions regarding space and time?

Any one of these questions can be dealt with in an honest way under restricted coverage of subject matter and can be used to cultivate and enhance aspects of scientific literacy such as those defined in Section 12.2.

By contrast, I suggest that enterprises such as the following (however popular they may be) are at best useless, and at worst damaging, since there is insufficient time to attack the “How do we know . . . ? Why do we believe . . . ?” questions or the necessary background is far too advanced:

- 1 Students being told about the “fascinating” particles of high-energy physics (with unintelligible jargon about interactions, angular momentum, mass-energy relations, quantum transitions, quarks, gluons, color, strangeness, the uncertainty principle) when they have inadequate understanding of concepts such as velocity, acceleration, force, mass, energy, and electrical charge, much less any understanding of how we obtain evidence regarding the structure of matter on a scale that transcends our direct sense perceptions.
- 2 Students who are still essentially Aristotelian, with no significant understanding of the law of inertia, satisfying distribution requirements by taking courses in meteorology or oceanography and hearing incomprehensible assertions about the role of the Coriolis effect.
- 3 Students who have no notion how to define local noon or the north-south direction, who have no idea of the origin of the seasons or of the phases of the moon (believing the unilluminated portion of the crescent moon to be the shadow of the earth), who are unaware that the stars have a diurnal motion, who do not understand why we believe the earth and planets revolve around the sun, taking “general education” courses in astronomy and hearing lectures on stellar nucleosynthesis, pulsars, quasars, and black holes.
- 4 Students who do not know how substances are defined and recognized, who have no idea what is meant operationally by words such as “oxygen,” “nitrogen,” or “carbon,” who do not understand why we believe in discreteness in the structure of matter, and who have no idea what is meant by “electrical charge” or “potential difference” being lectured to about DNA, molecular biology, and the structure of genes, or about nerve and muscle action.

The stream of unintelligible words cannot possibly generate scientific literacy; it simply aggravates the problem we are trying to solve.

12.5 ILLUSTRATING THE NATURE OF SCIENTIFIC THOUGHT

By addressing ideas such as those suggested in the preceding section—at a pace that allows formation and understanding of underlying concepts as well as consideration of the “How do we know . . . ? Why do we believe . . . ?” questions—illustrations of the character and limitations of scientific thought, rather than being injected artificially by assertion, will arise naturally and abundantly.

When, in the *Two New Sciences*, Galileo confronts the problem of describing the change in velocity of a moving body (the idea to which we give

the name “acceleration”), he points out that there are at least two alternatives: (1) we observe that the object changes its velocity from one value to another while it traverses a distance of so many cubits, and we might elect to describe this motion by means of the number that indicates how much the velocity changes for each successive cubit of displacement; (2) the same velocity change, however, occurs over a measurable interval of time, and one might also characterize the process by the number indicating how much the velocity changes in each successive second. Which mode of description should be adopted? The choice is not trivial.

Galileo selects the second concept: Change of velocity per unit time. His objective is to create a description of “naturally accelerated” motion (free fall), and he has a deep-seated intuitive conviction that free fall is uniformly accelerated in this sense but not in the sense of change of velocity per unit displacement. On the basis of a hypothesis, an inductive guess, he selects the concept that will yield the simplest and most elegant description of free fall and proceeds to test the hypothesis by deducing consequences that can be tested by experiment.

Here are, lucidly displayed, several significant facets of the scientific enterprise: The roles of inductive and deductive reasoning; the fact that scientific concepts are created by deliberate acts of human intelligence and imagination and are not “objects” discovered accidentally; that choice may be necessary and that there might be room for aesthetic criteria such as elegance and simplicity.

Here, too, is the new and revolutionary idea of forming an a priori hypothesis and testing its mathematical predictions by experiment. The Greeks had appealed on many occasions to both observation and experiment (they adduced, for example, the resistance of an inflated animal bladder to compression as experimental evidence for the corporeality of air), but they did not test the models or hypotheses they invented for the explanation of natural phenomena.

I outline this well-known story not to lay claim to new or profound insights into the philosophy or history of science, but only to point explicitly to a set of significant ideas that college students, with little prior scientific knowledge, can comprehend and appreciate—ideas that lie just below the surface in any introductory study of physical science, but that students are rarely given the time and opportunity to discover, articulate, and savor. To acquire these insights, students must have the opportunity to stand back and examine what happened, to relive some of the intellectual experience (the doubts as well as the successes), to analyze and assess the line of thought, recognizing the elements of its logic, its strength, and its limitations (what questions were not asked and not answered).

In only a few texts and courses, however, are students afforded this opportunity. The more common procedure is to throw the standard definitions of velocity and acceleration at the students in two or three cryptic pages (note the

difficulty most students have in forming these concepts clearly and correctly in the first place [Trowbridge and McDermott (1980), (1981)]. The concepts are asserted as though they were inevitable, rocklike formations that have existed for all time, while deference is paid to “history” by mention of the name of Galileo and by a few unsubstantiated clichés concerning his invention of the “experimental method” and his paternity with respect to “modern science.”

Students can indeed acquire mature and intelligent intellectual perspectives toward the methods, processes, successes, and limitations of science. Such perspectives, however, are not automatically conveyed by training students to calculate how high a stone rises when it is thrown vertically upward, or how much a given electrical field between capacitor plates will deflect an electron beam. Such intellectual perspectives can be developed only by coupling understanding of the scientific subject matter itself with the insights gained by standing back and examining not just the end results but what happened in terms of “How do we know . . . ? Why do we believe . . . ? What is the evidence for . . . ?” (See Section 13.2 in the discussion of critical thinking.)

Opportunities to make explicit such facets of the cultural phenomenon that is science arise at almost every turn without the necessity of invoking esoteric fringes of modern science that are completely unintelligible to students at their existing level of concept formation and grasp of subject matter. Consider a few additional examples.

In view of the rapid, assertive way in which scientific concepts are forced on students in school science as well as in introductory college courses, it is understandable that they acquire the notion that scientific terms are rigid, unchanging entities with only one absolute significance that the initiated automatically “know” and that the breathless student must acquire in one brain-twisting gulp. It comes as a revelation and a profound relief to many students to learn that scientific terms go through an evolutionary sequence of redefinition, sharpening, and refinement as one starts at a crude, initial, intuitive level, and, profiting from insights gained in successive applications, develops the concept to its subsequent level of sophistication.

For example, the concept of “force” is legitimately introduced by connection with the primitive, animate, intuitive, muscular sense of push or pull, but, in the law of inertia, we redefine it to apply to any interaction that imparts acceleration to a material object (for example, the action on bits of paper of a glass rod rubbed with silk.) We endow completely inanimate objects with the capacity to exert forces on other objects: The charged rod exerting a force on bits of paper, the table exerting an upward force on the book resting on it, the earth exerting a downward force on us (our “weight”), and the ground exerting an upward force at our feet. Following Newton, we then extend the concept even further and create the idea that, when the table exerts an upward force on the book, the book simultaneously exerts a downward force on the table, and, even more subtly, as the earth exerts a downward gravitational

force on us, we exert an oppositely directed gravitational force on the earth. By this time we have come a very long way indeed from the original use of the word “force” in connection with an animate, muscularly sensed push or pull on another object. And since the concept is very subtle indeed, even very able students have great difficulty assimilating it in the far too short time made available in introductory courses, while many less able students never assimilate it at all (see discussion of these matters in Chapter 3).

Similarly, starting with the crude idea of “speed” as a measure of how fast (as an average over a substantial time interval) an object travels along a straight line, we endow the concept with direction one way or the other along the straight line, and, with the introduction of plus and minus signs, we begin to call it “velocity.” We then refine this primitive notion into a conception of “instantaneous velocity” (something even the extremely able Greeks never achieved). We then extend the concept to cover direction in two- and three-dimensional space; and we finally talk about the rate of change of this quantity in both direction and magnitude.

At each stage in the sequence of evolution and redefinition, the original word (whether it be “force” or “velocity”)—its meaning having been changed in significant, intrinsic ways—no longer denotes only the first intuitive idea to which it was applied; it denotes a new and more sophisticated concept. (Strictly speaking, we ought to give it a new name in order to emphasize the change, but this, of course, would only enhance the verbal chaos.) Modest self-consciousness about the process of definition and redefinition enormously increases the confidence of students in their own grasp of the new sequence of thought, opens their eyes to similar shifts and extensions in the subsequent generation of concepts such as “energy,” “electrical charge,” “electron,” and carries over into other areas of study by alerting them to similar semantic shifts in courses in other disciplines—shifts, I might add, that are rarely explicitly pointed out or emphasized but that are crucial to genuine understanding. (In this connection, see discussion of the importance of the process of operational definition in Section 13.2.)

During the 1830s and 1840s, Michael Faraday’s beautiful and elegant experimental investigations of electricity and magnetism caused him to raise some very deep questions about these phenomena, and students can be led to articulate a few of these themselves if helped Socratically in group discussion: Is there a *process* by which one electrically charged particle exerts a force on another? If one of the particles is suddenly displaced, does a finite time interval elapse before the force experienced by the other changes? Does a finite time interval elapse between the instant a wire is connected to an electric battery and the instant a neighboring compass needle begins to swing in response to the magnetic effect of the electric current in the wire? If a finite time interval does elapse in each case, what, if anything, happens in the intervening space between the interacting objects? What happens to Newton’s third law and the whole concept of “action-at-a-distance” if the objects do not exert equal and

opposite forces on each other during the interval of change—however short it may be?

To try to address these questions, Faraday invented a “model,” a heuristic device completely transcending any direct sense experience—the famous concept of “lines of force,” “lines” that stretched, contracted, spread apart, and pulled together, propagating electric and magnetic effects through empty space—the concept that James Clerk Maxwell subsequently elaborated into the sophisticated modern, mathematical notion of “field.” Faraday wrote almost apologetically about his highly speculative model [Faraday (1965b)]:

It is not to be supposed for a moment that speculations of this kind are useless or necessarily hurtful in natural philosophy. They should ever be held as doubtful and liable to error and to change, but they are wonderful aids in the hands of the experimentalist and mathematician; for not only are they useful in rendering the vague idea more clear for the time, giving it something like a definite shape, that it may be submitted to experiment and calculation; but they lead on by deduction and correction, to the discovery of new phenomena, and so cause an increase and advance of real physical truth, which unlike the hypothesis that led to it, becomes fundamental knowledge not subject to change.

This fine description of the point and function of a heuristic model simultaneously reveals a characteristic facet of the thinking of many 19th century scientists: They were indeed convinced that they were stockpiling “real physical truth” and “knowledge not subject to change.” After students learn something of the conceptual revolution accompanying the failures of classical theory at the turn of the 20th century, it is interesting to ask them to contrast Faraday’s confident statement with the sadder and wiser one of J. Robert Oppenheimer (1963):

We come to our new problems full of old ideas and old words, not only the inevitable words of daily life, but those which experience has shown to be fruitful over the years. . . . We love the old words, the old imagery, and the old analogies, and we keep them for more and more unfamiliar and more and more unrecognizable things.

In light of such perspectives, students spontaneously begin to articulate some sense of why most scientists now view scientific knowledge as mutable and provisional rather than permanent and final. They anticipate limited ranges of validity to successful theories and are prepared to find a regression of unanswered questions behind every answered question.

In still another sequence involving models transcending direct sense experience, one can have students follow and examine the evidence that led to our

belief in atoms and molecules (discreteness in the structure of matter) as well as to belief in the structure of atoms themselves. They must be allowed to doubt with the early participants, to articulate uneasiness about interpretation of some of the evidence, and not just be stuffed with a few disconnected and, in themselves, unconvincing arguments, followed by assertion of the end results. (The original doubters were, after all, a far from foolish company.)

Many illustrative gems line the way through such a sequence. Dalton, for example, in his original attempts to develop a quantitative atomic-molecular theory with a unique and internally consistent set of relative atomic masses, confronted chemical data in the form of percentage composition by weight of various known compounds. The only regularity that had been discerned in these data was the so-called law of definite proportions—the recognition of a fixed percentage composition of any definite chemical substance regardless of its place or manner of origin—and even this aspect was a matter of some controversy and uncertainty. Dalton's preconceptions concerning the corpuscular constitution of matter, however, led him to give particular consideration to cases in which a given pair of elements (say carbon and oxygen) form more than one compound. It occurred to him that if 1.0 g of carbon combined with 1.3 g of oxygen in one compound, then, for the same 1.0 g of carbon in the other compound, one should find perhaps 2.6 or 0.65 or 3.9 g of oxygen—or some other quantity that bore a whole-number ratio to 1.3. One would expect just such simple numerical connections if compounds did indeed consist of molecules made up of small numbers of atoms of the combining elements. The data had never been examined in this way; this particular orderliness lay hidden behind the unrevealing percentage compositions.

Dalton looked and found that the predicted order was indeed there; in other words he predicted what is now called the “law of multiple proportions.” (In many modern courses, this law, if it is considered at all, is presented as though it had been established empirically along with the law of definite proportions and had provided *a priori* evidence for discreteness.) The point here, of course, is that very frequently “facts” do *not* speak for themselves. In this instance, the facts had been available for a long time, but they were not even seen until looked for through the lenses of a theory; then suddenly their uncovered presence fed back as a dramatic confirmation of the theoretical conception that had revealed them.

The story of the famous “Piltdown Man” fraud that was exposed in the 1950s is an inverse illustration of this idea. Many paleontologists accepted as genuine the fossil with the manlike skull and the apelike jaw because their theoretical preconceptions led them to expect an evolutionary sequence in which brain development led the way for changes in other parts of the body.

They accepted the forgery for almost 40 years, even though it was well known that it did not fit any reasonable niche in the humanoid fossil sequence. Here again, facts did not speak for themselves; they were viewed through the lenses of a theory, and the theory led many astray.

In a genuinely liberal (i.e., “liberating”) educational enterprise, it is essential that the broad, general characteristics of any intellectual enterprise be brought to light, but this cannot be accomplished through vague generalizations disconnected from the visceral effort that goes with forming concepts and understanding the subject matter. To develop scientific literacy, it is necessary to master at least some reasonable amount of subject matter to give meaning to the generalizations.

12.6 ILLUSTRATING CONNECTIONS TO INTELLECTUAL HISTORY

In reasonably paced introductory courses, there are many ways of making students aware of the role of science in their own intellectual history.

For example, in working in the field of elementary school science instruction, I have more than once encountered the following sequence: A child asks (or perhaps is asked by the teacher) “Why do objects fall?” and the “correct” response given or solicited by the teacher turns out to be “because of gravity.” The impression is instilled in the child that a reason has been given that explains both cause and effect. There is no inkling on the part of either the giver or the receiver of the “information” that the technical name conveys neither knowledge nor understanding and merely conceals ignorance of the nature of the phenomenon. If one asks a class of college students “Why do objects fall?” the great majority respond “because of gravity,” and few, if any, have the courage to say “we have no idea.”

Few students, teachers, or citizens have any awareness of the history of this term—that “gravity” started as the designation of a teleological effect, a “drive” or “desire” on the part of the “heavy” elements earth and water (and their mixtures) to seek the center of the earth; that the opposite “desire” on the part of air and fire to rise was called “levity”; that 17th century natural philosophy banished both the teleological view and the word “levity”; that Newton, explicitly eschewing knowledge of mechanism or process of interaction, made the grand surmise that, however it might work, the same effect that makes the apple fall binds the moon to the earth and the earth and planets to the sun; that, despite the beauty and elegance of the general theory of relativity, we have, to this day, no idea of how gravity “works.” In light of the answer most students give to the question “Why do objects fall?” a small step toward scientific literacy might reside in disabusing them of the foolish answer.

Very few students have any conception of the revolutionary thrust of 17th

century science in discarding the notion that the heavenly bodies were made of a “perfect” substance (the “quintessence”) intrinsically different from the four mundane materials of earth, and in discarding, also, the notion that the laws governing the celestial domain are entirely different from those of the terrestrial. The discarding of these notions together with the subsumption of the entire universe under one system of humanly comprehensible natural law, the extension of the universe to infinity, and the concomitant removal of the literal sheltering heaven from close overhead, marked a profound turning point in our intellectual history. The outlook of every individual toward himself and his place in the universe is deeply conditioned by this heritage from Galileo, Descartes, Newton and other 17th century natural philosophers. An educated person should be aware of such a heritage in conceptual, historical, and intellectual terms, not in the mere assertion of end results. Here is another significant step toward greater scientific literacy. Excellent presentations of this story, with both its scientific and intellectual aspects, are available at undergraduate levels [Holton (1973); Rutherford, Holton, and Watson (1981)], but they are not widely used even in general education courses, and they make virtually no appearance at all in curricula for scientists and engineers.

Yet if these historical and intellectual aspects were brought out explicitly in introductory science courses, they would enormously enhance and enrich any study of the Enlightenment in courses in Western Civilization and would stimulate a healthy diffusion of ideas through the artificial membranes that now divide one course from another. Students are, for example, astonished to find how many aspects of the rhetoric used by the Founding Fathers in our American historical documents are traceable through the Deists back to Locke and Newton [Arons (1975)].

In an entirely different domain, consider some examples from literature. In Shakespeare’s *Henry VI*, there is the passage:

Glory is like a circle in the water
Which never ceaseth to enlarge itself
Till by broad spreading it disperse to naught.

This figure is much appreciated by students who encounter it while studying and observing common wave phenomena.

In the “Morning Song of Senlin” by Conrad Aiken (1953), there occur the lines:

The earth revolves with me, yet makes no motion,
The stars pale silently in a coral sky.
In a whistling void I stand before my mirror,
Unconcerned, and tie my tie.

The first line alludes to the law of inertia and Galilean relativity; the “whistling void” refers to the Newtonian cosmology. And Aiken embellishes

all of this with a final touch of irony and paradox. This is clearly a modern view; nothing like it could possibly have been written in the 16th century. Here students can see a direct influence of science on literature, an episode in intellectual history affecting their own view of the universe in which they are placed, and they can do so, not in terms of unintelligible jargon, but in terms of scientific concepts developable and understandable at introductory level. (As a further step for those interested, it is illuminating to compare Aiken's matter-of-course acceptance of the Newtonian world with Milton's equivocation between the Ptolemaic and Copernican systems in *Paradise Lost*.)

Another tacit acceptance of the Newtonian universe emerges in Tennyson's lines in "Locksley Hall Sixty Years After":

While the silent heavens roll
 and suns upon their fiery way,
All their planets whirling round them,
 flash a million miles a day.

Another profoundly influential sequence, emanating from technology, overarching and eventually unifying all the sciences, and having a pervasive impact in intellectual history, is the story of the "principles of impotence." Beginning in early times with a growing recognition that nature conspires against our "getting something for nothing," against our achieving endless supplies of matter, motion, change, or warmth without effort and cost, the 18th century began to quantify the relevant restrictions. The "nothing can be created by divine power out of nothing" of Lucretius became, through the beautiful experiments of Lavoisier, the law of conservation of mass. The concepts of mechanical energy began to emerge, but their connection to thermal effects was not yet apparent.

The 19th century saw the grand synthesis of the first and second laws of thermodynamics: the unification of mechanical energy and heat; the conservation of energy overall; the possible and impossible directions of spontaneous change and the trend toward equilibrium; the possibility of converting a given amount of work entirely into thermal energy and the impossibility of converting a given amount of thermal energy entirely into work; the subsumption of all changes in state—mechanical, chemical, thermal, electric, magnetic—under just those two laws. The quantified forms of the principles of impotence, emerging out of the technology of the Industrial Revolution, brought impressive and comprehensible order out of the chaos of disparate forms of change, profoundly influencing the thought, for example, of social philosophers such as France's George Sorel [Humphrey (1951)] and of the American historian Henry Adams (1918).

Another sequence can follow a somewhat different direction. The 19th century had seen mass and energy as two separate, closed systems, each maintaining its own integrity of conservation. The 20th century, through the conceptual revolution precipitated by Einstein, began to comprehend mass and

energy as a unity rather than a dichotomy, as interconvertible as work and heat. This led to comprehension of how a radioactive material could maintain itself, apparently indefinitely, at a temperature higher than that of its surroundings; how the sun could keep shining for the enormous extent of geological time; and finally, how the prodigious energy confined in the atomic nucleus could be unleashed.

This sequence leads directly to all the moral, ethical, and societal problems that eventuate from the application of the resulting technology to peaceful industry on the one hand, and to warfare on the other. But the story and its influence on human society began long before Fermi and his associates constructed the first nuclear reactor in Chicago. I would hope to have students see some of this, with its ever-growing intensity, in the rich perspective of intellectual history, not just in a narrow view of the immediate scientific, technological, and societal end results.

12.7 VARIATIONS ON THE THEME

The preceding illustrations of epistemological, philosophical, and historical aspects of science, which, I submit, do cultivate some of the aspects of scientific literacy listed in Section 12.2 and do infuse liberal and humanistic perspectives into any curriculum, happen to be personal favorites of mine. I present them in order to be specific rather than vague in my illustrations and not to advocate them above the host of other possibilities. Each teacher must select subject matter and episodes that appeal to him or her and that he or she can articulate for the students in the most stimulating and compelling way.

There is the whole array of aspects to which James Conant referred as the “tactics and strategy” of science and which are effectively illustrated in the Harvard Case Histories in Experimental Science (1957). There is the fascinating, partly scientific, partly sociological problem of validation and acceptance of scientific theories. There are the philosophical problems of positivism and the questions concerning the “reality” of entities that transcend our senses: atoms, molecules, electrons, nuclei.

There are also, of course, the pressing social problems that stem from the release of nuclear energy; the application of science to warfare; the possibility of controlling human genetic development and even of synthesizing living matter; the problems of controlling and limiting abuse of our terrestrial environment. I have no intention of minimizing the significance of any of these vital questions, but I do have reservations about launching into analyses or discussions of them prior to the point at which students have some reasonable understanding of the underlying scientific and technical subject matter. If they have engaged in the necessary prior study, or if the necessary underpinnings are conscientiously developed as the need arises, such discussions can be educationally fruitful and significant and can enhance scientific literacy. If the questions are engaged without adequate understanding of the underlying

science, as is unfortunately frequently done, the enterprise becomes specious. Students are then misled into thinking that they have pursued inquiry and possess understanding when they have, in fact, only used technical terms the meaning of which they did not comprehend and only dealt in vacuous generalizations devoid of genuine thought and substance. In such circumstances, they have been subtly encouraged to embrace the all too widespread rationalization that “any opinion is just as good as any other opinion.”

It is clearly impossible to do all the things one would like to do with sequences such as those illustrated specifically in Sections 12.5 and 12.6 and mentioned briefly in this section. It is impossible to consider every important problem, open every significant perspective. The time is long past when we could teach our students all the things they need to know. It is hardly a new or original assertion that the only viable and realistic function of higher education is to put the students on their own intellectual feet: To give them conceptual starting points and an awareness of what it means to learn and understand something so that they may then continue to read, study, and learn, as need and opportunity arise, without perpetual formal instruction. (Such intellectual development is, of necessity, inextricably connected with the capacities for critical thinking discussed in the following chapter.)

An essential criterion is that students must not end up regurgitating secondhand pronouncements about the nature and processes of science without ever having articulated any such insights out of their own intellectual experience with subject matter they can encompass. Without at least some participation in comprehension and interpretation of scientific concepts, theories, and philosophy, students learn no more from secondhand statements about science than they learn from a commentary on poetry without having read the poetry, or a discussion on the methods and philosophy of history if they lack knowledge of the history of anything.

Rather than plunging into levels of material utilizing concepts and relationships well beyond the students’ present comprehension, if we were to allow them the opportunity to confront science through some of the more modest insights and experiences I have tried to illustrate above, I am convinced that they would develop lasting insights far more consonant with our postulated objectives of liberal education. Such insights would contribute to the development of better educated individuals and more thinking-reasoning citizens in exactly the same way as do awareness of history and sensibility to literature.

12.8 ASPECTS OF IMPLEMENTATION

Although I would encourage, especially at upper division levels, occasional courses devoted to specific important problems (such as those mentioned in Section 12.7), using team teaching—despite its relatively high cost—to assure interdisciplinary perspectives, I nevertheless believe that a great deal can be accomplished in the lower division level in existing curricular structure.

The walls separating courses in the sciences, the humanities, and the social sciences need to be made more permeable. In the illustrations I have given, I have tried to show how a science teacher might enhance the liberal content of courses by becoming, through personal effort, more literate and more articulate about historical, literary, philosophical, epistemological, humanistic, or societal overtones than the teacher was prepared to be in his or her own education. Materials to this end are readily available for both teachers and students, and only a little ingenuity is required to devise questions, homework problems, and paper assignments that lead students to confront ideas that transcend the mere end results constituting almost the entire burden of the great majority of our textbooks. Understanding what Galileo contributed to the development of modern science does not reside in being able to calculate how high the stone goes when we throw it upward; other ingredients must be included in addition to the calculation.

It is true that, in our present science courses, infusion of such perspectives would force us to back away from the already excessive volume and pace of “coverage” that are imposed on our willing and uncritical students. It has for long been my contention that we are crushing our students into the flatness of equation-grinding automatons and forcing them into blind memorization of problem-solving procedures. We do not even give them a chance to begin to understand what “understanding” really means. Were we to require that they understand concepts and lines of reasoning within the perspectives I have outlined, we would be setting far higher intellectual standards and demanding far higher performance than we now attain, with all our obsession with “coverage.” Some of our colleagues complain that this would mean a “watering down” of our science teaching, but they misunderstand what is being advocated. What I am suggesting would be far more a “raising up” than a “watering down.” An essential ingredient would be conscientious attention to the problems of teaching, learning, reasoning, concept formation, and understanding discussed in the earlier chapters of this book. To the development of secure command of concepts and reasoning, we must add some of the intellectual perspectives defined in the last few sections.

The intellectual perspectives, incidentally, should not be confined to the general education courses for nonscience majors. Future scientists and engineers are just as much in need of capacity for scientific literacy as are the nonscientists, and such literacy is not cultivated in our existing science and engineering courses. Solving the end-of-chapter problems, however correctly, does not open historical, philosophical, or humanistic perspectives and does not automatically cultivate the insights underpinning genuine scientific literacy. The most compelling evidence of our failure in enhancing the scientific literacy of our scientists and engineers resides in the widespread reluctance of college and university faculty members to inject and articulate humanistic ideas in their own courses. Most of them ran away from humanities courses because they were “not very good in humanities,” and they never saw any

parallel ideas made relevant in science. The fear of ideas about science (as opposed to the hard content itself) is, in many instances, more a matter of insecurity because of a faulty education rather than just stubborn insistence on the necessity of coverage.

I know at first hand that students can and do respond to the demand both for thorough understanding of the science and for penetration of the intellectual perspectives. It is futile, however, to exhort the students regarding the importance of liberal education and to give them nothing but ringing platitudes. We must be sufficiently knowledgeable and concerned to bring before them compelling and significant instances of intellectual, moral, or aesthetic experience. To this end, we, as instructors in science or engineering, must exhibit our own capacity to deal with such dimensions, else the exhortation will be hypocritical. How can we demand that the students do something we will not do ourselves?

But the burden is not on the technical instructors alone. The walls must be made permeable in the other direction also. Instructors in the humanities and social sciences must cease running from "hard science," as most of them did during their undergraduate careers; they must bring themselves to master at least a few of the "How do we know . . . ? Why do we believe . . . ?" questions mentioned earlier, and they must treat scientific subject matter and its consequences knowledgeably in their own courses—for example, Newton's influence should be seriously discussed (not casually and superficially) in any history course dealing with the 18th century enlightenment. The humanist and the social scientist must set an example for the students in the same way I demand of my technical and scientific colleagues. Again, how can we expect the students to cease running from anything that makes them feel slightly insecure if we persist in running ourselves?

As indicated repeatedly, to achieve these objectives we must cut back on the volume and pace of coverage that have escalated in *all* our courses. Students must have time to form concepts, think, reason, and perceive relationships. They must discuss ideas, and they must write about them. Examples of reduced coverage and selection of reasonable story lines do exist in our textbook literature, although the examples are relatively few in number. To illustrate what I mean, rather than advocating these texts specifically, I point to *The Project Physics Course* as an example of how to weave a story line that brings one to the beginnings of modern atomic and subatomic physics, leaving out unnecessary subjects along the way and introducing many of the intellectual perspectives advocated in this essay; I also point to Casper and Noer's (1972) *Revolutions in Physics* as a sequence that deals with Newton and Einstein without trying to cover everything else in between. Building on existing good examples, it is quite possible for an instructor to generate a reasonable sequence that best fits his or her own skills and predilections.

12.9 THE PROBLEM OF COGNITIVE DEVELOPMENT

If we accept the premise that our liberal education objectives entail thinking, reasoning, and understanding on the part of the students, it is essential that we take into account the empirical results of recent studies of cognitive development. Such studies [Arons (1976); Chiappetta (1976); Karplus et al. (1979); McKinnon and Renner (1971); Perry (1970)], usually based on administration of two or more of the now classic Piagetian tasks, are showing, with remarkable reproducibility, that only about 25% of the cross section of college students execute these tasks successfully, indicating that they have developed the capacity for abstract logical reasoning at this rudimentary level; that up to 50% of the students are unsuccessful in the tasks and still use predominantly concrete patterns of reasoning, while the remaining 25% are in transition, exhibiting only partial success.

Students do have intellects capable of development beyond this point. Such development, however, is necessarily based on repetitive practice and experience. Most college students are very much in need of practice in the various modes of thinking and reasoning listed and discussed in Chapter 13.

As has been pointed out earlier, the volume and pace of material thrust on students in the majority of liberal education science courses preclude the exercise of the time-consuming operations of thinking, reasoning, and understanding. The majority of students are thus forced into blind memorization, and they eventually come to see all “knowledge” and “understanding” as the juxtaposition of memorized names and phrases. They are not held to, or tested on, the reasoning since their teachers, pandering to student pressures, accede that it is “too hard.” Most students are tested almost exclusively on the memorized end results. In such circumstances, there is essentially no hope of developing insights and understanding that characterize genuine scientific literacy—or, for that matter, literacy in any field requiring abstract logical reasoning.

If we really wish to succeed in our liberal education objectives, it is essential to start from the premise that the students do have intellects capable of development and use. Then it is necessary to give them the opportunity to think, to reason, and to develop intellectually by providing material at a pace and level that make assimilation possible. Genuine understanding of a limited range of significant subject matter would testify to far higher intellectual standards and requirements than regurgitation of memorized jargon from advanced or topical subjects, however tantalizing the latter might be.

12.10 THE PROBLEM OF TEACHER EDUCATION

It is not difficult to see that far and away the greatest leverage and highest potential for improving public understanding of the subject matter, methods, limitations, and social impact of science reside in the elementary and secondary

schools. This is not to say that sophisticated, college-level insights should be developed in early schooling—pupils at that stage are not ready for such discussion; the point is that the groundwork can be laid for the synthesis of adult insight when maturity has been attained.

With genuine understanding of basic concepts that can perfectly well be developed in elementary and secondary school, with concomitant enhancement of capacities for abstract logical reasoning, college and university levels of instruction would be enormously facilitated. Students would enter with levels of knowledge and understanding that would make discussion of philosophical, historical, ethical, and societal questions fruitful and meaningful. They would even be able to penetrate aspects of modern science that are now hopelessly unintelligible. With such improved background, even our mass production system of lecturing to large classes might be substantially more effective, if still not ideal.

Such progress is impeded at the present time, not by lack of adequate curricular materials at elementary and secondary levels, but by inadequate teacher education in the sciences. It is unlikely that any curricular materials, however high their potential for cultivating thinking, reasoning, learning, and understanding, will ever be “teacher proof.” A teacher can always negate the intent of the materials by attitude, invidious comment, and, most significantly, by what he or she chooses to test for. The real intellectual values of a course are established by tests rather than by texts. If improved curricular materials are to be successfully implemented, we must have teachers who are secure in the use of the instructional materials. “Security” in this context means both thorough understanding of the subject matter and thorough understanding of the underlying pedagogical intent and design of the materials. The latter understanding is impossible without the former.

For example, intrinsic weakness of the instructional materials is not the cause of the too modest success in implementing existing fine, hands-on elementary school curricula [e.g., *Elementary Science Study* (1968), or *Science Curriculum Improvement Study* (1968)]. Rather, it is that college science preparation fails to provide elementary teachers with the knowledge and understanding necessary for effective use and implementation. Our work at the University of Washington with both pre- and in-service elementary teachers shows that the two groups are initially indistinguishable: The majority use concrete rather than formal patterns of reasoning; the majority cannot do arithmetical reasoning involving ratios and division (i.e., very few can solve word problems such as those invoked in fifth and sixth grade arithmetic); the majority fail on tasks involving control of variables; the majority cannot visualize, in the abstract, possible or plausible outcomes of changes imposed on a system. Their “knowledge” of science resides exclusively in memorized names, phrases, and technical terms, and, because they lack operational understanding of these terms, they are unable to reason with them in any specific instance.

In other words, our elementary teachers, if they had any physical science at all, have had courses of the variety criticized in Section 12.4. It is quite apparent that such courses do not help the future teacher exercise patterns of abstract logical reasoning, and they do not engender understanding of the subject matter. The future teacher is left with a half-remembered string of words such as “nuclei,” “laser,” “strange particles,” “angular momentum,” “black hole.” The teacher is left, furthermore, with no understanding of the concept of density and its relation to floating and sinking, of the law of inertia and its relation to why we believe the earth and planets revolve around the sun, of the distinction between heat and temperature, of the origin of the phases of the moon, of the observed apparent annual motion of the sun, of the definition of “north-south” or of “noon-midnight,” of the concept of electric current and its behavior in the very simplest resistive circuit, of the most elementary aspects of wave motion, of the most prevalent everyday phenomena of light.

The better elementary science curricula wisely pursue the realms of everyday experience rather than esoteric vocabularies of modern physics. The teachers find themselves at exactly the same level of initial knowledge as the children, and, because of consequent lack of security, they cannot handle the instructional materials. I have heard honest and perceptive teachers volunteer this assessment of their present condition in exactly these terms.

Time and again in our investigations, we have confirmed the empirical observation that adults (pre- and in-service elementary teachers), when they have not had the prior learning experiences as children, acquire the abstract reasoning patterns and the conceptual understanding fostered in the elementary science curricula only by encountering and overcoming the same obstacles, hurdles, and difficulties that are experienced by the children. And the fact of being adults is of virtually no help; the pace of learning is usually slower than it is among the children.

A major part of the effort expended in short workshops intended to help schoolteachers implement the new elementary science curricula was completely wasted because of the failure to recognize this necessary time element. Their college background has left the vast majority of teachers at the same point as the pupils they are expected to teach. It is illusory to hope that a brief indoctrination about the “philosophy of the program,” followed by rapid examination of a few sample units of the materials, will induce understanding of the content; it is also illusory to hope that, once they get “started,” the teachers will learn from the beautifully transparent materials along with the children. We have found that only slow, patient traversal of both the subject matter and the patterns of abstract logical reasoning gives the teacher the level of security that leads to effective implementation [Arons (1977)]. In the meantime, unless and until we change our college courses and curricula [McDermott (1990)], we will continue to produce elementary, intermediate, and secondary school teachers who are in need of remediation the instant they

graduate. This, however, is not inevitable; investigation in our group has shown that, given the opportunity to start at the beginning with hands-on materials and proceeding at a pace that allows learning and understanding, the great majority of pre- and in-service teachers do develop the requisite reasoning capacities and subject matter knowledge, and take great satisfaction in having done so. What they have lacked is only the opportunity.

A very common manifestation of the current inadequacy of the subject matter background of elementary school teachers is their reaction to the better hands-on curricular materials that are available. In some instances where the materials have been purchased by the schools, they stay locked up in closets, while the teachers, if they are obliged to teach science, proceed to lecture out of the standard older texts. In other instances, teachers, insecure in the face of the new materials, finding them "too difficult" for the children without being aware that the trouble really lies in their own lack of adequate understanding, band together and direct their energies and good intentions to writing materials of their own. The result is invariably trash that is full of errors, misconceptions, and misstatements, and that probably has negative educational value. (It must be reemphasized that the existing excellent elementary school materials were developed by mature and highly competent scientists who had the necessary perspectives and understanding.)

Research on teaching and learning may give us better instructional materials, but these materials will never have a really widespread impact in the schools unless we produce teachers who can implement them. This is a major problem, and it remains rooted at the college-university level. Accusations and recriminations leveled against elementary and secondary school teachers are, for the most part, unwarranted and unjust: "The fault, dear Brutus, is not in our stars"

Besides understanding of subject matter and enhancement of capacity for abstract logical reasoning, one other aspect of teacher education that deserves attention is behavior in the classroom. The best new science curricula are organized not around lecturing and inculcation by the teacher, but around exploration, trying and erring, talking and listening, arguing and explaining, among the children themselves. (There is evidence that, when competently implemented, such materials simultaneously enhance reading comprehension and arithmetical reasoning capacities in the children [Bredderman (1982); Shymansky et al. (1983); Wellman (1978)]. In many instances, when children are following an erroneous course, it is important that a teacher direct attention to inconsistencies or contradictions and guide them to revision or correction without flat assertion of "the right answer." Such teaching behavior takes confidence, security, and firm command of concepts and lines of reasoning. Lecturing and asserting the "right answer" are much easier and are the refuge of the insecure.

But even with adequate subject matter background, it is asking too much to expect such sophisticated behavior of a young person thrown into a new

classroom. The preservice teacher needs both prior example and instruction. Many researchers in this country and abroad consistently show that teachers tend to teach as they have been taught. If we do nothing but lecture at our preservice teachers, they will lecture in their own classrooms, regardless of the indoctrination that may have been given (also by lecturing) in their “methods” courses in departments of education. It is essential that we set aside courses for future teachers taught in small classes in exactly the way we wish to have the elementary science curricula implemented. The science course then becomes a methods course as well—as it should. With teacher education of this kind, we might begin to hope for a shift in performance in the schools, a shift that has not taken place through the provision of improved curricular materials alone. Such a shift would tremendously enhance the possibility of cultivating adult scientific literacy at college-university level.

12.11 A ROLE FOR THE COMPUTER

One outcome of research and observation over a wide range of students and introductory courses is that many students do not break through to full command of a particular concept or line of reasoning unless they can be reached in one-on-one Socratic dialogue. The breakthrough is not made in homework or exercises or tests or by passive listening to explanations, however lucid. A gradual breakthrough begins only after the learner has begun to articulate ideas, inferences, and lines of reasoning in his or her own words [Arons (1976), (1977), (1982)]. This is especially true of pre- and in-service elementary and secondary teachers.

Examples of such areas of difficulty are arithmetical reasoning using ratios and division; translating symbols into words and words into symbols (i.e., interpreting graphs of any kind, e.g., interpreting position-time and velocity-time graphs in kinematics both in words and by executing the indicated motion with one’s hand; translating word problems into arithmetical form; translating an arithmetical or algebraic expression into words, etc.); distinguishing operationally between “heat” and “temperature”; forming and using the model for the origin of phases of the moon (i.e., being able to deal successfully with questions such as “Would you expect to see a full moon rising at midnight? Why or why not?”); forming and using the model of current and the concept of electrical resistance to predict brightness changes that would result from alterations (shorts, removal or insertion of other bulbs) imposed on simple circuits composed of batteries and flashlight bulbs [Arons (1982)].

The necessary one-on-one dialog with a single student can easily take as long as 20 to 30 minutes or more. It is barely possible for an instructor to do this in small classes, and it is quite out of the question in large ones. One might adopt the view that students who cannot master such simple and basic concepts or reasoning modes do not belong in science courses. Enlightened self-interest, however, if not a broader objective, dictates a less callous

view. Among the students who fail or who simply disappear from our college-university courses (or who never enroll in the first place because of deep-seated fear and insecurity), and who could be saved by one-on-one help at strategic points, are many potentially promising minority students as well as most of our future elementary school teachers, not to speak of many others.

Here is a point at which modern technology can be of assistance. The coming of age of the personal computer with graphic capability offers the prospect of making one-on-one dialogues practicable in spite of numbers [Arons (1984d), (1986)]. The problem becomes one of writing effective dialogues that pull students over the early, most severe obstacles, and help them on the way to further learning, with decreasing dependence on Socratic assistance.

The writing of effective dialogues will of necessity lean heavily on the results of research in teaching and learning and in cognitive development. Not only must a skillful author have a broad awareness of what is being learned in general about cognitive development and utilization of patterns of abstract logical reasoning, but he or she must also draw heavily on the detailed concept-by-concept student protocols, that is, researches such as those discussed in earlier chapters of this book and listed in the bibliography. Many of us incline to make conjectures concerning what students were thinking and misapprehending in various circumstances. In making these conjectures, we are invariably extrapolating from our own personal experience and preconceptions, and such conjectures are, in fact, very rarely correct. The actual difficulty is almost always something plausible to the student that we have glossed over and have not perceived.

Yet for a really effective computer dialogue, the most important (and most difficult) provisions an author must make are the ones that lead a student to rectify incorrect responses; it is easy to take care of the correct ones. Socratic rectification of misconceptions and incorrect reasoning can be achieved only if the author has prior knowledge as to the actual incorrect responses likely to be made. This is why authors must be well versed in the research results if they are to write good material.

12.12 LEARNING FROM PAST EXPERIENCE

With present resurgence of long-standing concern about education of scientists and engineers and about public understanding of science and technology, some individuals are pointing to what they regard as the "failure" of the curriculum development efforts of the '50s and '60s, and are advocating a new round of preparation of "up-to-date" text materials, hoping that this time the materials will be "better" and "more effective." Without incorporation, however, of the additional ingredients and insights discussed in this book, this hope is fallacious and forlorn. Although existing curricular materials can certainly be improved in pedagogical quality, this alone will not overcome the real obstacles

and underlying problems. Curricular materials, however lucid, skillful, and imaginative, cannot “teach themselves” [cf Arons (1981a)].

I do not consider the science curriculum development of the '50s and '60s to have been a failure. At the time these efforts were undertaken, principally under stimulus provided by the National Science Foundation, new text materials, especially ones rooted in laboratory and observational experience, were very badly needed. Existing texts were obsolete, full of errors and misstatements, and intellectually sterile. They had been copied and recopied from each other for several generations by authors who themselves did not have adequate understanding of the subject matter.

In the heady, evangelistic atmosphere of 40 years ago, leaders in the various scientific and technical fields brought imaginative insight, current perspectives, and correct logical underpinnings to the new materials. It is true that the materials were not of uniformly high instructional quality. Some were too rapidly paced for the age groups for which they were intended. Some started by asserting end results of sophisticated lines of inquiry, required students to memorize these end results without examination of underlying “How do we know . . . ? Why do we believe . . . ?” questions, and proceeded to drown the students in incomprehensible consequences of the initial, unsupported assertions. Some of the curriculum developers, not having the benefit of more recent insights into the actual levels of capacity of different age groups for abstract logical reasoning [Arons (1976); Chiappetta (1976); Karplus et al. (1979); McKinnon and Renner (1971)], went off into abstractions completely beyond the grasp of the pupils they were ostensibly addressing (and of many teachers as well). Mathematics, in particular, suffered from such misdirection.

However, despite some misplaced effort and mismatch to intended audience, a significant number of excellent curricula were developed. In my opinion, the best are among those designed for elementary school [e.g., *Elementary Science Study* (1968); *Science Curriculum Improvement Study* (1968)]. In these materials children are started from scratch, with no presuppositions concerning prior “knowledge” of technical terminology. Ideas come first and names afterwards. Concepts are synthesized out of observational experience rather than received through lecture and assertion. Reasoning starts at concrete levels and, provided it is guided by a competent teacher, gradually proceeds toward the abstract. Repeated intellectual experiences of this kind stimulate the emergence of abstract reasoning capacity and enhance its subsequent development. Some sound and potentially effective materials were also developed for junior high school and high school levels. In many instances, however, these materials are well matched only to the upper levels of student development and are mismatched to a large proportion of the audience the developers had in mind. Although these materials are, in some instances, well suited to students one, two, or even three years older, they are rarely used at the higher age levels because of the inhibition imposed by the nominal lower age label.

The point is that, regardless of some deficiencies and mismatch, a sub-

stantial body of pedagogically sound and useful material was developed. Few of these materials are obsolete or “out of date” even after 20 or more years. Their adoption and implementation has indeed been disappointing, and some critics who wish to energize a new burst of curriculum development refer to them as a failure. The “failure” that is being adduced, however, does not stem from deficiencies in the quality of the materials so much as it stems from external causes that have, so far, not been remedied. Although the better of the existing curricula are surely not the last word in educational sophistication, and although improvements are certainly possible in light of improving insights into aspects of teaching and learning, I am convinced that a new generation of materials would suffer exactly the same “failure” in adoption and implementation from the same extrinsic causes.

There were two principal causes of such failure. One was inadequacy of logistic support, especially in elementary and junior high school, for teachers venturesome enough to try the new curricula. Hands-on, laboratory-oriented materials require continual maintenance and resupply. Busy and overloaded teachers need sustained support in using such materials; they cannot take care of the logistics in addition to all the other duties they are expected to discharge. Although they might make the extra effort in the initial wave of enthusiasm (and many of them did), such effort is impossible to sustain over long periods of time. The necessary logistic support was rarely strong, even at the very beginning, and it rapidly melted away altogether in the seventies with growing financial pressures on the schools.

Thus, one formidable obstacle that must be overcome at all levels in the schools is that of logistic support. It is illusory to suppose that widespread scientific literacy will ever be successfully cultivated through instructional materials based on purely verbal inculcation. The necessary understanding, reasoning, and mastery of basic concepts and ideas will evolve, in the great majority of ordinary individuals, only from concrete observational experience. It is true that some gifted individuals do break through to scientific understanding, and to abstract reasoning in general, without such concrete help. But this small fraction of the population has always broken through, perhaps with some delay, simply because of the availability of books, classes, homework, and explanation; they have probably done so in spite of, rather than because of, the instructional system. We are indeed fortunate that this is the case, for otherwise literacy of every variety would be far lower than it actually is. Our reiterated goal, however, is a large increase in the literate fraction. This goal will not be attained without providing the majority of our young students with hands-on experience and with sustained guidance in carrying concept formation and reasoning from the concrete to the abstract.

The second, and probably most significant, cause of failure in adoption and implementation of the better curricular materials (including those in social science as well as natural science) lies in what happened in the retraining of teachers. Huge amounts of money and effort were poured into summer insti-

tutes and academic year programs, most of them under the direction of college and university faculty. A few of these programs were beneficial because the staff recognized that the teachers had developed little genuine understanding of scientific concepts and subject matter in their previous school and college courses and were very nearly at the same level of conceptual development as the children they were supposed to teach. Accordingly, it was realized that the teachers must be guided slowly and carefully through the same intellectual experiences they were subsequently to convey in their own classrooms. In the majority of teacher training programs, however, this necessity was not understood or appreciated. It was falsely assumed that the teachers, elementary through high school level, really understood the very elementary science they were to be teaching, or that, if understanding was deficient, they would quickly develop it along with the pupils, simply by working with the perfectly lucid new materials.

In the summer institutes, and in other programs, the teachers were thus given lectures about the "educational philosophy" of materials the substance of which they did not comprehend. Or they were given more advanced subject matter, also totally incomprehensible, in order to extend their "perspectives" through awareness of up-to-date knowledge and progress in various fields. Or in the largest number of cases, they were simply given another run through the same excessively rapid, irrelevant, and unintelligible college courses that had had no visible intellectual effect in the past. In such instances, the teachers were no better able to handle the new curricular materials competently than they were before coming into the retraining programs. It was not their fault that they had no awareness of this deficiency and that, in subsequent, disappointing classroom experiences with the new materials with which they were still insecure, they ascribed their troubles to the difficulty of the materials rather than to their own inadequate understanding. (Let me hasten to say that I am not tarring all teachers with this brush. I know personally, and I continue to meet and hear about, superlative, competent teachers at every level. These are, however, far too small a fraction of the total teacher population to make for a solution to the national problem.)

If we wish to produce widespread improvement in public understanding of science, we will have to take significant steps toward mitigating the two principal causes of our previous failure. The problems are very difficult, but they are not intrinsically insoluble. The remedies are necessarily fairly costly, but by far the most difficult part of the problem is conveying comprehension of it to our college and university faculty colleagues, most of whom still operate on the premise that instruction of future teachers (as well as all other students) can be effectively conducted by sufficiently lucid verbal inculcation and through the range of subject matter "coverage" that has become conventional in introductory science courses. If they continue in this practice, we shall make no progress regardless of expenditure of effort on curriculum improvement.

Curricular materials, however fine, will never achieve our goals entirely by

themselves. They may improve matters for the especially gifted, but they will not solve the problem for the vast and perfectly respectable majority. Unless the essential underpinnings I have been trying to define are provided, tinkering with new curricula will again absorb huge amounts of time and money and will leave us exactly where we are at the present time. We will have a new crop of complaints about failure and inadequacy of the materials and a revived demand for still another needless wave of curriculum development. Attainment of wider scientific literacy will keep receding into the infinite educational perspective.

Chapter 13

Critical Thinking

The simple but difficult arts of paying attention, copying accurately, following an argument, detecting an ambiguity or a false inference, testing guesses by summoning up contrary instances, organizing one's time and one's thought for study—all these arts . . . cannot be taught in the air but only through the difficulties of a defined subject; they cannot be taught in one course in one year, but must be acquired gradually in dozens of connections.

JACQUES BARZUN

13.1 INTRODUCTION¹

No curricular recommendation, reform, or proposed structure has ever been made without some obeisance to the generic term “critical thinking” or one of its synonyms. The flood of reports on education in our schools and colleges that has been unleashed in recent years is no exception; every report, at every level of education, calls for attention to the enhancement of thinking-reasoning capacities in the young. A currently prominent formula is “higher order thinking skills.” Few of the documents that come to us, however, attempt to supply some degree of specificity—some operational definition of the concept, with illustrations of what might be done in day-to-day teaching to move toward the enunciated goals.

It is the object of this chapter to try to “unpack” the term “critical thinking”—to list a few simpler, underlying processes of abstract logical reasoning that are common to many disciplines and that can be cultivated and exercised separately in limited contexts accessible to the student. Subsequently, the individual’s conscious weaving together of these various modes results in the larger synthesis we might characterize as “critical thought.” As Barzun points out in the quotation cited above, this can be done only through practice in, preferably, more than one field of subject matter.

¹This chapter is based on an article originally published in *Liberal Education* [Arons (1985)].

13.2 A LIST OF PROCESSES

To glimpse some of the ways in which effective schooling might enhance students' reasoning capacities, it is instructive to examine a few of the thinking and reasoning processes that underlie analysis and inquiry. These are processes that teachers rarely articulate or point out to students; yet these processes are implicit in many different studies. The following listing is meant to be illustrative; it is neither exhaustive nor prescriptive. Readers are invited to add or elaborate items they have identified for themselves or sense to be more immediately relevant in their own disciplines.

1 Consciously raising the questions "What do we know . . . ? How do we know . . . ? Why do we accept or believe . . . ? What is the evidence for . . . ?" when studying some body of material or approaching a problem.

Consider the assertion, which virtually every student and adult will make, that the moon shines by reflected sunlight. How many people are able to describe the simple evidence, available to anyone who can see, that leads to this conclusion (which was, incidentally, perfectly clear to the ancients)? This does not require esoteric intellectual skills; young children can follow and understand; all one need do is lead them to watch the locations of both the sun and moon, not just the moon alone, as a few days go by. Yet for the majority of our population the "fact" that the moon shines by reflected sunlight is received knowledge, not sustained by understanding.

Exactly the same must be said about the contention that the earth and planets revolve around the sun. The validation and acceptance of this view marked a major turning point in our intellectual history and in our collective view of man's place in the universe. Although the basis on which this view is held is more subtle and complex than that for the illumination of the moon, the "How do we know . . . ?" should be an intrinsic part of general education; it is, for most people, however, received knowledge—as is also the view that matter is discrete in its structure rather than continuous.

Similar questions should be asked and addressed in other disciplines: How does the historian come to know how the Egyptians, or Babylonians, or Athenians lived? On what basis does the text make these assertions concerning consequences of the revocation of the Edict of Nantes? What is the evidence for the claim that such and such tax and monetary policies promote economic stability? What was the basis for acceptance of the doctrine of separation of church and state in our political system?

Cognitive development researchers [e.g., Anderson (1980); Lawson (1982)] describe two principal classes of knowledge: figurative or declarative on the one hand, and operative or procedural on the other. Declarative knowledge consists of knowing "facts" (matter is composed of atoms and molecules; animals breathe oxygen and expel carbon dioxide; the United States entered the Second World War after the Japanese attack on Pearl Harbor in Decem-

ber 1941). Operative knowledge involves understanding where the declarative knowledge comes from or what underlies it (What is the evidence that the structure of matter is discrete rather than continuous? What do we mean by the terms “oxygen” and “carbon dioxide” and how do we recognize these as different substances? What worldwide political and economic events underlay the American declaration of war?). And operative knowledge also involves the capacity to use, apply, transform, or recognize the relevance of declarative knowledge in new situations.

“Above all things,” says Alfred North Whitehead in a well-known passage on the first page of *The Aims of Education*, “we must beware of what I will call ‘inert ideas’—that is to say, ideas that are merely received into the mind without being utilized, or tested, or thrown into fresh combinations.” And John Gardner once deplored our tendency to “to hand our students the cut flowers while forbidding them to see the growing plants.”

Preschool children almost always ask “How do we know . . . ? Why do we believe . . . ?” questions until formal education teaches them not to. Most high school and college students then have to be pushed, pulled, and cajoled into posing and examining such questions; they do not do so spontaneously. Rather, our usual pace of assignments and methods of testing all too frequently drive students into memorizing end results, rendering each development inert. Yet given time and encouragement, the habit of inquiry can be cultivated, the skill enhanced, and the satisfaction of understanding conveyed. The effect would be far more pronounced and development far more rapid if this demand were made deliberately and simultaneously in science, humanities, history, and social science courses rather than being left to occur sporadically, if at all, in one course or discipline.

2 Being clearly and explicitly aware of gaps in available information. Recognizing when a conclusion is reached or a decision made in absence of complete information and being able to tolerate the attendant ambiguity and uncertainty. Recognizing when one is taking something on faith without having examined the “How do we know . . . ? Why do we believe. . . ?” questions.

Interesting investigations of cognitive skill and maturity are conducted by administering test questions or problems in which some necessary datum or bit of information has been deliberately omitted, and the question cannot be answered without securing the added information or making some plausible assumption that closes the gap. Most students and many mature adults perform very feebly on these tests. They have had little practice in such analytical thinking and fail to recognize, on their own, that information is missing. If they are told that this is the case, some will identify the gap on reexamining the problem, but many will still fail to make the specific identification.

In our subject matter courses, regardless of how carefully we try to examine evidence and validate our models and concepts, it will occasionally be

necessary to ask students to take something on faith. This is a perfectly reasonable thing to do, but it should never be done without making students aware of what evidence is lacking and exactly what they are taking on faith. Without such care, they do not establish a frame of reference from which to judge their level of knowledge, and they fail to discriminate clearly those instances in which evidence has been provided from those in which it has not.

3 Discriminating between observation and inference, between established fact and subsequent conjecture.

Many students have great trouble making such discriminations even when the situation seems patently obvious to the teacher. They are unused to keeping track of the logical sequence, and they are frequently confused by technical jargon they have previously been exposed to but never clearly understood.

In the case of the source of illumination of the moon cited earlier, for example, students must be made explicitly conscious of the fact that they see the extent of illumination increasing steadily as the angular separation between moon and sun increases, up to full illumination at a separation of 180° . This direct observation leads, in turn, to the inference that what we are seeing is reflected sunlight.

In working up to the concept of "oxygen" (without any prior mention of this term at all) with a group of elementary school teachers some years ago, I had them do an experiment in which they heated red, metallic copper in an open crucible and weighed the crucible periodically. What they saw happening, of course, was the copper turning black and the weight of crucible and contents steadily increasing. When I walked around the laboratory and asked what they had observed so far, many answered, "We observed oxygen combining with the copper." When I quizzically inquired whether that was what they had actually seen happening, their reaction was one of puzzlement. It took a sequence of Socratic questioning to lead them to state what they had actually seen and to discern the inference that something from the air must be joining the copper to make the increasing amount of black material in the crucible. It had to be brought out explicitly that this "something from the air" was the substance to which we would eventually give the name "oxygen." What they wanted to do was to use the technical jargon they had acquired previously without having formed an awareness of what justified it.

This episode illustrates the importance of exposing students to repeated opportunity to discriminate between observation and inference. One remedial encounter in one subject matter context is not nearly enough, but opportunities are available at almost every turn. Mendel's observations of nearly integral ratios of population members having different color and size characteristics must be separated from inference of the existence of discrete elements controlling inheritance. In the study of literature, analysis of the structure of a novel or a poem must be distinguished from an interpretation of the work.

In the study of history, primary historical data or information cited by the historian must be separated from the historian's interpretation of the data.

A powerful exercise once employed by some of my colleagues in history was to give the students a copy of the Code of Hammurabi accompanied by the assignment: "Write a short paper addressing the following question: From this code of laws, what can you infer about how these people lived and what they held to be of value?" This exercise obviously combines exposure to both processes 1 and 3.

4 Recognizing that words are symbols for ideas and not the ideas themselves. Recognizing the necessity of using only words of prior definition, rooted in shared experience, in forming a new definition and in avoiding being misled by technical jargon.

From the didactic manner in which concepts (particularly scientific concepts) are forced on students in early schooling, it is little wonder that they acquire almost no sense of the process of operational definition and that they come to view concepts as rigid, unchanging entities with only one absolute significance that the initiated automatically "know" and that the breathless student must acquire in one intuitive gulp. It comes as a revelation and a profound relief to many students when they are allowed to see that concepts evolve; that they go through a sequence of redefinition, sharpening, and refinement; that one starts at crude, initial, intuitive levels and, profiting from insights gained in successive applications, develops the concept to final sophistication.

In my own courses, I indicate from the first day that we will operate under the precept "idea first and name afterwards" and that scientific terms acquire meaning only through the description of shared experience in words of prior definition. When students try to exhibit erudition (or take refuge from questioning) by name dropping technical terms that have not yet been defined, I and my staff go completely blank and uncomprehending. Students catch on to this game quite quickly. They cease name dropping and begin to recognize, on their own, when they do not understand the meaning of a term. Then they start drifting in to tell us of instances in which they got into trouble in a psychology, or sociology, or economics, or political science course by asking for operational meaning of technical terms. It is interesting that this is an aspect of cognitive development to which many students break through relatively quickly and easily. Unfortunately, this is not true of most other modes of abstract logical reasoning.

5 Probing for assumptions (particularly the implicit, unarticulated assumptions) behind a line of reasoning.

In science courses, this is relatively easy to do. Idealizations, approximations, and simplifications lie close to the surface and are quite clearly articulated in most presentations. They are ignored or overlooked by the students, however, principally because explicit recognition and restatement are rarely, if

ever, called for on tests or examinations. In history, humanities, and the social sciences, underlying assumptions are frequently more subtle and less clearly articulated; probing for them requires careful and self-conscious attention on the part of instructors and students.

6 Drawing inferences from data, observations, or other evidence and recognizing when firm inferences cannot be drawn. This subsumes a number of processes such as elementary syllogistic reasoning (e.g., dealing with basic propositional, “if . . . then” statements), correlational reasoning, recognizing when relevant variables have or have not been controlled.

Separate from the analysis of another’s line of reasoning is the formulation of one’s own. “If . . . then” reasoning from data or information must be undertaken without prompting from an external “authority.” One must be able to discern possible cause-and-effect relations in the face of statistical scatter and uncertainty. One must be aware that failure to control a significant variable vitiates the possibility of inferring a cause-and-effect relation. One must be able to discern when two alternative models, explanations, or interpretations are equally valid and cannot be discriminated on logical grounds alone.

As an illustration of the latter situation, I present a case I encounter very frequently in my own teaching. When students in a general education science course begin to respond to assignments leading them to watch events in the sky (diurnal changes in rising, setting, and elevation of the sun, waxing and waning of the moon, behavior of the stars and readily visible planets), they immediately expect these naked eye observations to allow them to “see” the “truth” they have received from authority, namely that the earth and planets revolve around the sun. When they first confront the fact that both the geo- and heliocentric models rationalize the observations equally well and that it is impossible to eliminate one in favor of the other on logical grounds at this level of observation, they are quite incredulous. They are shocked by the realization that either model might be selected provisionally on the basis of convenience, or of aesthetic or religious predilection. In their past experience, there has always been a pat answer. They have never been led to stand back and recognize that one must sometimes defer, either temporarily or permanently, to unresolvable alternatives. They have never had to wait patiently until sufficient information and evidence were accumulated to develop an answer to an important question; the answer has always been asserted (for the sake of “closure”) whether the evidence was at hand or not, and the ability to discriminate decidability versus undecidability has never evolved.

An essentially parallel situation arises in the early stages of formation of the concepts of static electricity (see Sections 6.7 and 6.8). Students are very reluctant to accept the fact that, before we know anything about the microscopic constitution of matter and the role of electrical charge at that level, it is impossible to tell from observable (macroscopic) phenomena whether

positive charge, negative charge, or both charges are mobile or being displaced. They wish to be told the “right answer” and fail to comprehend that any one of the three models accounts equally well for what we have observed and predicts equally well in new situations. They want to use the term “electron” even though they have no idea what it means or what evidence justifies it, and they apply it incorrectly to irrelevant and inappropriate situations.

If attention is explicitly given, experiences such as the ones just outlined can play a powerful role in opening student minds to spontaneous assessment of what they know and what they do not know, of what can be inferred at a given juncture and what cannot.

7 Performing hypothetico-deductive reasoning; that is, given a particular situation, applying relevant knowledge of principles and constraints and visualizing, in the abstract, the plausible outcomes that might result from various changes one can imagine to be imposed on the system.

Opportunities for such thinking abound in almost every course. Yet students are most frequently given very circumscribed questions that do not open the door to more imaginative hypothetico-deductive reasoning. The restricted situations are important and provide necessary exercises as starting points, but they should be followed by questions that impel the student to invent possible changes and pursue the plausible consequences.

8 Discriminating between inductive and deductive reasoning; that is, being aware when an argument is being made from the particular to the general or from the general to the particular.

The concepts of “electric circuit,” “electric current,” and “resistance” can be induced from very simple observations made with electric batteries and arrangements of flashlight bulbs. This leads to the *inductive* construction of a “model” of operation of an electric circuit. The model then forms the basis for *deductive* reasoning, that is, predictions of what will happen to brightness of bulbs in new configurations or when changes (such as short circuiting) are imposed on an existing configuration.

Exactly similar thinking can be developed in connection with economic models or processes. Hypothetico-deductive reasoning is intimately involved in virtually all such instances, but one should always be fully conscious of the distinction between the inductive and the deductive modes.

9 Testing one’s own line of reasoning and conclusions for internal consistency and thus developing intellectual self-reliance.

The time is long past when we could teach our students all they need to know. The principal function of education—higher education in particular—must be to help individuals to their own intellectual feet: To give them conceptual starting points and an awareness of what it means to learn and understand something so that they can continue to read, study, and learn as need and opportunity arise, without perpetual formal instruction.

To continue genuine learning on one's own (not just accumulating facts) requires the capacity to judge when understanding has been achieved and to draw conclusions and make inferences from acquired knowledge. Inferring, in turn, entails testing one's own thinking, and the results of such thinking, for correctness or at least for internal coherence and consistency. This is, of course, a very sophisticated level of intellectual activity, and students must first be made aware of the process and its importance. Then they need practice and help.

In science courses, they should be required to test and verify results and conclusions by checking that the results make sense in extreme or special cases that can be reasoned out simply and directly. They should be led to solve a problem in alternative ways when that is possible. Such thinking should be conducted in both quantitative and qualitative situations. In the humanities and social sciences, the checks for internal consistency are more subtle, but they are equally important and should be cultivated explicitly. Students should be helped to sense when they can be confident of the soundness, consistency, or plausibility of their own reasoning so that they can consciously dispense with the teacher and cease relying on someone else for the "right answer."

10 Developing self-consciousness concerning one's own thinking and reasoning processes.

This is perhaps the highest and most sophisticated reasoning skill, presupposing the others that have been listed. It involves standing back and recognizing the processes one is using, deliberately invoking those most appropriate to the given circumstances, and providing the basis for conscious transfer of reasoning methods from familiar to unfamiliar contexts.

Given such awareness, one can begin to penetrate new situations by asking oneself probing questions and constructing answers. Starting with artificial, idealized, oversimplified versions of the problem, one can gradually penetrate to more realistic and complex versions. In an important sense, this is the mechanism underlying independent research and investigation.

13.3 WHY BOTHER WITH CRITICAL THINKING?

The preceding list of thinking and reasoning processes underlying the broad generic term "critical thinking" is neither complete nor exhaustive. For illustrative purposes, I have tried to isolate and describe processes and levels of awareness that appear to be bound up with clear thinking and genuine understanding in a wide variety of disciplines and to show a deep commonality in this respect among very different kinds of subject matter. These processes underlie the capacity defined by Jacques Barzun in the quotation that heads this chapter.

Developing these intellectual skills requires extensive, sustained practice.

Such practice is not possible in a space devoid of subject matter. It is only through contact with, and immersion in, rich areas of subject matter that interesting and significant experience can be generated. Although it may be possible, in principle, to generate limited aspects of such practice through artificial kinds of exercises and puzzle solving, or even through analysis of scores in sports contests, it seems a waste of time to resort to such sterile channels when all the vital disciplines of our culture lie at our disposal.

Why should we want to cultivate skills such as those I have listed? There are many obvious reasons having to do with quality of life, with professional competence, with the advance of culture and of society in general, but I particularly wish to suggest a socio-political reason: the education of an enlightened democratic citizenry. What capacities characterize such a citizenry?

Justice Learned Hand, the distinguished jurist of the preceding generation, argued with telling irony that we would be able to preserve civil liberties only so long as we were willing to engage in the "intolerable labor thought, that most distasteful of all our activities." John Dewey in *Democracy and Education* contends that "The opposite to thoughtful action are routine or capricious behavior. Both refuse to acknowledge responsibility for the future consequences which flow from present action."

The requirements set by Barzun, Hand, and Dewey can be broken down to more fundamental components. The sophisticated distinction between enlightened and short range self-interest is based on hypothetico-deductive reasoning. Such reasoning is also inevitably involved in visualizing possible outcomes of decisions and policies in economic and political domains.

There is need to discriminate between facts and inferences in the contentions with which one is surrounded. There is the necessity of making tentative judgments or decisions, and it is better that this be done in full awareness of gaps in available information than in an illusion of certainty. There is the highly desirable capacity to ask critical, probing, fruitful questions concerning situations in which one has little or no expertise. There is the need to be explicitly conscious of the limits of one's own knowledge and understanding on a given issue.

Each of these capacities appears on the preceding list, and I believe that each can be cultivated and enhanced, at least to some degree, in the great majority of college students through properly designed experiences embracing a wide variety of subjects.

I hasten to emphasize that these skills alone are not sufficient to assure good citizenship or other desirable qualities of mind and person. Other ingredients are necessary, not the least of which are moral and ethical values, which impose their own constraints on the naked processes of thinking and reasoning. Although values are not disconnected from thinking and reasoning, the educational problems they pose transcend the limits of this short essay and require discussion in their own right.

13.4 EXISTING LEVEL OF CAPACITY FOR ABSTRACT LOGICAL REASONING

In the United States some investigators have rather belatedly come to realize that much of our science curricular material, and the volume and pace with which we thrust it at our students, are badly mismatched to the existing levels of student intellectual development at virtually every age. I am convinced that the same is true in other disciplines, but the fact is less readily discerned because assignments and tests concentrate on end results and procedures rather than on reasoning and understanding.

I say that “some” have become aware of this problem because, despite the unequivocal and relentlessly accumulating statistics, many who teach in the schools, colleges, and universities remain unaware of the emerging data; others fail to see any relevance to their own teaching.

Beginning about 1971, investigators began administering elementary tasks in abstract logical reasoning (such as those pioneered by Jean Piaget [see Piaget and Inhelder (1958)] in his studies of the development of abstract reasoning capacity in children) to adolescents and adults of college age and beyond [see, for example, Chiapetta (1976); McKinnon and Renner (1971)]. The tests have centered principally on arithmetical reasoning with ratios or division and on awareness of the necessity of controlling variables in deducing cause-effect relationship.

Although the results vary significantly from one population to another (economically disadvantaged versus economically advantaged; concentrating in science and engineering versus concentrating in humanities or fine arts versus concentrating in the social sciences, etc.), the overall averages have remained essentially unchanged with increasing volume of data since the first small samples were reported in 1971, and, most suggestively, the averages do not change appreciably with increasing age beyond about 12 or 13: Roughly one third of the total number of individuals tested solve the tasks correctly; roughly one third perform incorrectly but show a partial, incipient grasp of the necessary mode of reasoning; the remaining third fail completely. In Piagetian terminology, the first group might be described as using formal patterns of reasoning, the third group as using principally concrete patterns, and the middle group as being in transition between the two modes [Arons and Karplus (1976)].

The weaknesses revealed by these two specific tasks would mean relatively little if they stood by themselves, but, in fact, these weaknesses are closely correlated with weaknesses in other modes of abstract logical reasoning such as discriminating between observation and inference; dealing with elementary syllogisms involving inclusion, exclusion, and serial ordering; recognizing gaps in available information; doing almost any kind of hypothetico-deductive reasoning.

Most of the curricular materials thrust at students in the majority of their courses at secondary and college level implicitly require well-developed rea-

soning capacity in the modes that have been listed in this discussion. In fact, only a small proportion of the students (less than one third) are ready for such performance. The rest, lacking the steady, supportive help and explicit exercises required, resort, in desperation, to memorization of end results and procedures. Failing to develop the processes underlying critical thinking, they fail to have experience of genuine understanding and come to believe that knowledge is inculcated by teachers and consists of recognizing juxtapositions of arcane vocabulary on multiple choice tests. (Readers familiar with the studies of William G. Perry will recognize his first category of intellectual outlook among college student [Perry (1970)].)

13.5 CAN CAPACITY FOR ABSTRACT LOGICAL REASONING BE ENHANCED?

In our Physics Education Group at the University of Washington, we have worked intensively for some years with populations of pre- and in-service elementary school teachers and other nonscience majors ranging in age from 18 to over 30. Initially no more than about 10% were using formal patterns of reasoning. By starting with very basic, concrete observations and experiences, forming concepts out of such direct experience, going slowly, allowing students to make and rectify mistakes by confronting contradiction or inconsistency, insisting that they speak and write out their lines of reasoning and explanation, repeating the same modes of reasoning in new contexts days and weeks apart, we have been able to increase the fraction who successfully use abstract patterns of reasoning to perhaps 70 to 90%, depending on the nature of the task.

The most important practical lesson we have learned is that repetition is absolutely essential—not treading water in the same context until “mastery” is attained, but in altered and increasingly richer context, with encounters spread out over time. Quick, remedial exercises in artificial situations preceding “real” course work are virtually useless. One must patiently construct repeated encounters with the same modes of reasoning in regular course work and allow students to benefit from their mistakes. Progress becomes clearly visible in the sense that the percentage of successful students increases with each repetition.²

²On any one kind of task (e. g., arithmetical reasoning involving division, or forming a clear operational and intuitive distinction between mass and volume), the following history of progress is typical: For the first few repetitions the percentage of successful performers increases substantially, but the curve is concave downward and invariably levels off somewhere between 70 and 90% after four to six opportunities [Arons (1976); Rosenquist (1982)]. We have never been able to achieve 100% success. Approximately 15% of the students never developed, under our guidance, the capacity to perform successfully on the given task, even with further repetition and with intensive personal tutoring.

There are certain obvious questions: Were we insufficiently skillful in providing guidance

It is still a very long step from the development of specific abstract reasoning processes in one area of subject matter, such as elementary science, to more advanced levels of subject matter in the same area, not to speak of transfer to entirely different areas. What little evidence exists suggests that very little transfer occurs from experience acquired in only one discipline. I myself am strongly convinced, however (mostly by fragmentary, anecdotal evidence, and perhaps some admixture of wishful thinking), that very great progress could be effected if students were simultaneously exposed to such intellectual experience in entirely different disciplines. This is largely a matter of conjecture since an organized experiment at the college level has not really been tried.

The fragmentary evidence to which I appeal comes from two disparate sources:

1 Experience with a tightly organized core curriculum at Amherst College during the '50s and '60s: In this curriculum, there was a very strong interaction among an English composition course, a science course, an American Studies course, and, toward the later stages, a Western Civilization course, all of which had certain attitudes, approaches, and intellectual standards in common [Kennedy (1955); Arons (1978)]. Alumni of that period tend to comment very favorably, in retrospect, on the effect of that experience on their own intellectual development. (So tightly organized a curriculum was a rather special case and, as then implemented, would be possible only with a small, homogeneous student body. Judicious modifications should, however, be effective in more heterogeneous situations.)

2 Data being reported on effects of the elementary school science curricula developed under auspices of the National Science Foundation during the '60s. The latter evidence is very indirect, but it is highly suggestive and merits a bit of discussion. The groups that developed the new curricula worked directly with the children they sought to teach and met the latter on their own ground and at their existing verbal and conceptual starting points rather than in some never-never land of unchecked and untested hypotheses and assumptions about children and learning. Everything in these materials begins with hands-on experience and observation. Concepts are developed through induction and synthesis from this experience, with the teacher as guide and pilot rather than as verbal inculcator. Ideas are developed *first* and names are invented *afterwards*; technical terms are generated operationally only after experience has given them sanction and meaning. [As examples, the reader might refer to

and instruction to the unsuccessful 15%? Would success still be achieved over much longer periods of time? What would have happened if these individuals had received such instruction at the age of 11 or 12 instead of so much later? Are there some individuals who are intrinsically unable to develop these capacities? Are our observations, for some reason, invalid? All we can say is that we do not know the answers. We had reached a point of diminishing returns under available time and resources and were unable to press the issue further. The empirical fact is that the progress curve leveled off below 100%. We hope that answers to some of the preceding questions will begin to emerge as time goes by.

programs such as *Elementary Science Study* (1968) and *Science Curriculum Improvement Study* (1968) both available from Delta Education, Nashua NH.]

The essence of instruction in these programs, whether the subject matter is physical or biological, is to give the children time—time to explore, to test, to manipulate, to talk and argue about meaning and interpretation, to articulate hypotheses, to follow trails to dead ends and retrace steps if necessary, to make mistakes and to revise views and interpretations when guided to perceive contradictions (instead of being told, by assertion, that their idea was “right” or “wrong”), to decide when and how arithmetical calculations should be made.

Such learning is sometimes (misleadingly) called “discovery learning.” The children are, of course, not expected to be Newtons, Faradays, Agassizes, or Darwins, “discovering” the concepts and theories of science *de novo* by the age of ten. Ordinary, lively, curious children simply react positively to the opportunity to learn from perceptively guided experience and observation. They retain what they learn because they are synthesizing genuine experience rather than memorizing a jumble of meaningless and unfamiliar words. They know where their knowledge comes from and are able to address the “How do we know . . . ? Why do we believe . . . ?” questions.

Since the first appearance of these curricula, researchers have been comparing the achievement of children exposed to such materials with the achievement of controls. In addition to showing significantly improved command of science subject matter, children exposed to the new curricula show significantly greater progress in both verbal and numerical skills, and the effects are particularly strong among disadvantaged children [cf. Bredderman (1982); Shymansky, et al. (1983); Wellman (1978)]. In other words, this mode of instruction, when competently implemented, results in transfer, enhancing performance beyond the science subject matter alone.

Although there is no direct evidence of a similar kind supporting the notion that we would enhance the higher level reasoning skills of college students by undertaking the instructional effort I have been advocating, I submit that the observations of the effect of the inquiry-oriented science curricula on children are at least very encouraging. The processes involved are analogous, and the effort seems worth making.

13.6 CONSEQUENCES OF MISMATCH

We are indeed fortunate that a significant proportion of our student population, perhaps one quarter, *does* make the breakthrough on modes of abstract logical reasoning spontaneously. (Consider the consequences to our society if this were otherwise!) But this does not lessen the urgency of improving our performance. As pointed out earlier, there now exists a serious mismatch between curricular materials and expectations on the one hand and actual

level of student intellectual development on the other. The curricular materials implicitly require abstract reasoning capacities and levels of insight and interpretation that many students have not yet attained. Neither the materials nor the most prevalent modes of instruction provide the gently paced, insistent, repetitive guidance that is necessary for helping students develop the necessary intellectual skills.

This mismatch has extensive deleterious consequences. We force a large fraction of students into blind memorization by imposing on them, particularly at high school and university levels, materials requiring abstract reasoning capacities they have not yet attained. And we proceed through these materials at a pace that precludes effective learning and understanding, even if the necessary reasoning capacities have been formed. Under such pressure, students acquire no experience of what understanding really entails. They cannot test their “knowledge” for plausible consequences or for internal consistency; they have no sense of where accepted ideas or results come from, how they are validated, or why they are to be accepted or believed. In other words, they do not have the opportunity to develop the habits of critical thinking defined earlier in this essay, and they acquire the misapprehension that knowledge resides in memorized assertions, esoteric technical terminology, and regurgitation of received “facts.” Although such failure is widely prevalent in the sciences, it is by no means confined there. It pervades our entire system, including history, the humanities, and the social sciences.

One specific example of the mechanism through which an entire system becomes degraded emerges through our experience with arithmetical reasoning. When I first discovered that no more than 10% of my undergraduate nonscience majors could reason arithmetically with division, I wondered what had happened to the old word problems that were used to cultivate such reasoning from the fifth and sixth grades on. Going back to existing elementary school arithmetic texts, I found that such problems were still there, as in my own school days, and were probably significantly improved. When I questioned my university students, they began to reveal that they had never actually had to do such problems in school because the problems were “too hard.” When I began working with in-service elementary school teachers and found that they themselves could not deal with such problems, the pattern was clear: An engineer would describe the system as a “degenerative feedback loop.” The arithmetical reasoning disability of the future teachers had never been detected and remedied when they were at the university. They graduated, went into the schools, and passed their disability and fear to most of the children by not requiring the doing of the word problems and conveying the rationalization that they were “too hard.” The children went on to the university, and so on, and so forth.

The case of arithmetical reasoning is just an especially clear and vivid illustration. The same pattern arises over and over again in other instances: In failure to master and understand the most fundamental scientific concepts

(such as velocity and acceleration or the nature of floating and sinking); in poor writing and speaking of English; in incapacity to deal with historical reasoning and the concomitant blind concentration of historical “facts.”

I wish to emphasize most strongly that the teachers whose incapacities I describe are not the ones to be blamed for this situation. The input terminals to the feedback loop of my metaphor reside in *our* hands at the colleges and universities. *We* are the ones who perpetuate the mismatch and fail to provide remediation of disabilities and enhancement of abstract reasoning capacities at the opportunities that we control. *We* are the ones who made the teachers as they are.

The mismatch about which I complain affects, of course, not only our future teachers but the whole of our student population outside the 25% who, in spite of the system, manage to break through spontaneously to abstract reasoning patterns. I dwell so insistently on the teachers only because of the crucial role they play in sustaining the feedback loop. Think of the prodigious impetus that might stem from altering the condition of the teachers and making the feedback regenerative instead of degenerative! What might we be able to achieve at university level if the mismatch between our materials and student readiness were removed?

13.7 ASCERTAINING STUDENT DIFFICULTIES

It might be helpful to point out some hard facts regarding the securing of reliable information concerning student learning difficulties and levels of abstract reasoning. What one must learn to do is ask simple, sequential questions, leading students in a deliberate Socratic fashion. After each question, one must shut up and listen carefully to the response. (It is the tendency of most inexperienced questioners to provide an answer, or to change the question, if a response is not forthcoming within one second. One must learn to wait as long as four or five seconds, and one then finds that students, having been given a chance to think, will respond in sentences and truly reveal their lines of thought.)

As the students respond to such careful questioning, one can begin to discern the errors, misconceptions, and missteps in logic that are prevalent. One learns nothing by giving students “right answers” or “lucid explanations.” As a matter of fact, students do not benefit from such answers or explanations; they simply memorize them. Students are much more significantly helped when they are led to confront contradictions and inconsistencies in what they say and then spontaneously alter their own statements as a result of such confrontation.

In such dialogs, two things immediately strike novice investigators. First they find that virtually all of their a priori conjectures concerning what students are and are not thinking are incorrect and that entirely unanticipated

but very fundamental, plausible, and deeply rooted preconceptions, misconceptions, and misapprehensions (of which the investigator had no awareness) are revealed. Second, they discover the saving grace in all of this unanticipated complexity: The frequently voiced cliché that every individual is completely different from every other individual is patently untrue. Each kind of misconception or erroneous mode of reasoning occurs, with remarkable reproducibility, in many individuals. Some hurdles and misconceptions are very widely prevalent. When one finds an approach or insight that overcomes a particular difficulty, that approach will be helpful not to only one but to many individuals.

It must be strongly emphasized that conclusions must be based on careful and accurate listening to students. Casual extrapolation of one's own experience only leads to error. Those of us who are fortunate enough to have become competent professionals are among the 25% minority mentioned earlier. We made the breakthrough in spite of the system, not because of it. Our own learning experiences are not representative, and citing such experience rarely leads to correct insight into what transpires for the majority of learners.

13.8 TESTING

At present, deficiency in the quality of testing is one of the more serious ills of our profession. There is a large and perceptive literature on testing in virtually every discipline, but its influence has, unfortunately, not been extensive. Some cynics have even remarked on the existence of a destructive collusion between students and teachers—a collusion in which students agree to accept bad teaching provided they are given bad examinations.

It is useless to render lip service to sophisticated intellectual goals and then test only for end results, vocabulary, “facts,” or “information.” The real goals of a course are determined not by what we say but what we test for. Students quickly ascertain what the real requirements of a course are and orient their efforts accordingly. Their attention can be focussed on the higher intellectual processes and requirements only if these aspects are included in testing and writing and play an important role in the final grade.

It is my earnest hope that more self-conscious attention to thinking and reasoning processes on the part of faculty will lead to statistical improvement in the quality of test questions and writing assignments. Good questions are very hard to devise, and any one individual runs out of inspiration. Collaborative effort could greatly increase the pool of good material and also provide the debugging that is always necessary.

13.9 SOME THOUGHTS ON FACULTY DEVELOPMENT

Given the almost universally accepted goal of enhancing the capacity for critical thinking in our students, it seems reasonable to lead faculty members

to sharpen their own critical thinking about how this goal is to be attained through the use of units of subject matter in their own areas of expertise—units which they have taught and with which they feel comfortable. The problem is to get away from vague, mushy generalizations and to provide constraints that induce consideration and elaboration of very specific examples: analyzing a unit of subject matter so as to identify the thinking and reasoning processes that must be brought to bear by the student, and devising questions that lead the student to such penetration.

I suggest that useful results might stem from the organization of faculty workshops in which participants, working in pairs, come prepared with a response to something like the following assignment:

- (a) Select from within your area of expertise a unit of subject matter (or a laboratory experiment) with which you are thoroughly familiar and study of which, you believe, will help a student attain a particular intellectual goal or insight or that will serve to exercise a particular reasoning process. (The unit should be as short as possible, but it should have a significant goal and not end up as a triviality.)
- (b) State the goals or insights that you discern.
- (c) Describe the essence of the unit of subject matter; that is, indicate how it provides a path toward the goal or insight.
- (d) List the various abstract reasoning capacities that the student must already possess, or must be helped to develop, in order to deal properly with the subject matter and not have to resort to memorization. (The basis for this analysis might be the list provided earlier in this chapter or an appropriately modified or augmented list.)
- (e) Indicate the kind of help you might provide students who encounter difficulty penetrating the material (e.g., questions that help point up significant issues, or clarify concepts, or focus on assumptions that are likely to be overlooked).
- (f) Indicate what writing you might ask the students to do in connection with the unit and how you would test for final mastery or understanding.
- (g) Indicate how you would lead the students to stand back, become conscious of the patterns of thinking and reasoning in which they had engaged, and, if possible, connect this experience with experiences they have had in other courses.

I imagine the workshop as bringing together pairs of individuals from the same and different disciplines. Each pair would present their analysis for discussion and comment by the entire group. There need be no “expert” or “authority” directing the proceedings. I would like to think that faculty

members seriously interested in the intellectual development of their students would find such an exercise interesting and stimulating. They would become more conscious of commonalities across disciplinary lines while defining real differences more precisely. The necessity of presenting the essence of a specific intellectual exercise to colleagues in other disciplines would help minimize the use of jargon and would sharpen awareness of how units of subject matter can be utilized. And finally, the whole enterprise would help cultivate an instructional climate in which students clearly perceived that they were being helped to develop their own capacities for critical thinking.

Part I Bibliography

- ADAMS, H. (1918), *The Education of Henry Adams* (Houghton Mifflin, Boston (many subsequent editions)).
- ADLER, C. (1958), "On the Humanization of Some Physics Problems," *Am. J. Phys.* **26**, 42.
- AIKEN, C. (1953), *Collected Poems* (Oxford University Press, New York).
- ANDERSON, J. R. (1980), *Cognitive Psychology and Its Implications* (Freeman, San Francisco).
- ANDERSON, J. R. ed. (1981), *Cognitive Skills and Their Acquisition* (Lawrence Erlbaum Associates, Hillsdale, NJ).
- ARONS, A. B. (1965), *Development of Concepts of Physics* (Addison-Wesley, Reading, MA).
- ARONS, A. B. (1975), "Newton and the American Political Tradition," *Am. J. Phys.* **43**, 209.
- ARONS, A. B. (1976), "Cultivating the Capacity for Formal Reasoning," *Am. J. Phys.* **44**, 834.
- ARONS, A. B. (1977), *The Various Language: An Inquiry Approach to the Physical Sciences* (Oxford University Press, New York).
- ARONS, A. B. (1978), "Teaching Science" in *Scholars Who Teach*, CAHN, S. M., ed. (Nelson-Hall, Chicago).
- ARONS, A. B. (1979), "Basic Physics of the Semidiurnal Lunar Tide," *Am. J. Phys.* **47**, 934.
- ARONS, A. B. (1981a), "Whither Do We Hurry Hence?" in *AAPT Pathways*. Proceedings of the Fiftieth Anniversary Symposium of the AAPT. (Published by the American Association of Physics Teachers.)
- ARONS, A. B. (1981b), "Thinking, Reasoning and Understanding in Introductory Physics Courses," *The Physics Teacher* **19**, 166.
- ARONS, A. B. (1982), "Phenomenology and Logical Reasoning in Introductory Physics Courses," *Am. J. Phys.* **50**, 13.
- ARONS, A. B. (1983a), "Achieving Wider Scientific Literacy," *Daedalus* Spring 1983.
- ARONS, A. B. (1983b), "Student Patterns of Thinking and Reasoning, Part One," *The Physics Teacher* **21**, 576.
- ARONS, A. B. (1984a,b), "Student Patterns of Thinking and Reasoning, Parts Two and Three," *The Physics Teacher* **22**, 21, 88.
- ARONS, A. B. (1984c), "Education Through Science," *J. Col. Sci. Teaching* **13**, 210.

- ARONS, A. B. (1984d), "Computer-Based Instructional Dialogs in Science Courses," *Science* **224**, 1051.
- ARONS, A. B. (1985), "Critical Thinking and the Baccalaureate Curriculum," *Liberal Education* **71**, 141.
- ARONS, A. B. (1986), "Overcoming Conceptual Difficulty in Physical Science Through Computer-Based Socratic Dialogs," in *Designing Computer-Based Learning Materials*, WEINSTOCK, H., and BORK, A., eds. Proceedings of the NATO workshop held in San Miniato, Italy, July 1985. NATO ASI Series F: *Computer and Systems Sciences*, Vol. 23 (Springer-Verlag, Berlin, New York).
- ARONS, A. B. (1989), "Developing the Energy Concepts in Introductory Physics," *The Physics Teacher* **27**, 506-517.
- ARONS, A. B. (1993), "Guiding Insight and Inquiry in Introductory Physics Laboratory," *The Physics Teacher* **31**, 278-282.
- ARONS, A. B., and BORK, A. M. (1964), "Newton's Laws of Motion and the 17th Century Laws of Impact," *Am. J. Phys.* **32**, 313.
- ARONS, A. B., and KARPLUS, R. (1976), "Implications of Accumulating Data on Levels of Intellectual Development," *Am. J. Phys.* **44**, 396.
- ARONS, A. B., and PEPPARD, M. B. (1965), "Einstein's Proposal of the Photon Concept: A Translation of the *Annalen der Physik* Paper of 1905," *Am. J. Phys.* **33**, 367.
- BARTLETT, A. A. (1976-1979), "The Exponential Function" (Parts I-IX) *The Physics Teacher* **14**, 393, 485; **15**, 37, 98, 225; **16**, 23, 92; **17**, 23.
- BIRKS, J. B. (1962), *Rutherford at Manchester* (Heywood & Co. Ltd., London).
- BOHR, N. (1913), "On the Constitution of Atoms and Molecules," *Phil. Mag.* **26**, 6, 1.
- BORK, A. (1979), "Interactive Learning," *Am. J. Phys.* **47**, 5.
- BORK, A. (1981), *Learning With Computers* (Digital Press, Bedford, MA).
- BRASELL, H. (1987), "The Effect of Real-Time Laboratory Graphing on Learning Graphic Representations of Distance and Velocity," *J. Res. Sci. Teach.* **24**, 385.
- BREDDERMAN, T. (1982), "Effects of Activity-Based Science in Elementary School," in *Education in the '80s: Science*. ROWE, M. B., ed. (National Education Association, Washington, DC).
- BRIDGMAN, P. W. (1941), *The Nature of Thermodynamics* (Harvard University Press, Cambridge, MA).
- BRIDGMAN, P. W. (1962), *A Sophisticate's Primer of Relativity* (Wesleyan University Press, Middletown, CT. Revised edition 1983).
- BROWN, S. C. (1958), "Do College Students Benefit From High School Laboratory Courses?" *Am. J. Phys.* **26**, 334.
- BRUSH, S. G. (1961), "John James Waterston and the Kinetic Theory of Gases," *Am. Scientist* **49**, 202.
- BRUSH, S. G. (1965), *Kinetic Theory. Vol. 1. The Nature of Gases and of Heat* (Pergamon Press, Oxford).
- CASPER, B. M., and NOER, R. J. (1972), *Revolutions in Physics* (W. W. Norton & Co., New York).
- CHABAY, R., and SHERWOOD, B. (1995) *Electric and Magnetic Interactions* (John Wiley & Sons, Inc., New York).
- CHAMPAGNE, A. B., GUNSTONE, R. F., and KLOPPER, L. E. (1985), "Instructional Consequences of Students' Knowledge about Physical Phenomena," in

- Cognitive Structure and Conceptual Change*, WEST, L. H. T., and PINES, A. L., eds. (Academic Press, Orlando, FL).
- CHAMPAGNE, A. B., KLOPPER, L. E., and ANDERSON, J. H. (1980), "Factors Influencing the Learning of Classical Mechanics," *Am. J. Phys.* **48**, 1074.
- CHIAPPETTA, E. L. (1976), "A Review of Piagetian Studies Relevant to Science Instruction at the Secondary and College Level," *Sci. Ed.* **60**, 253.
- CLEMENT, J. (1979), "Mapping a Student's Causal Conceptions from a Problem Solving Protocol," in *Cognitive Process Instruction*, LOCHHEAD, J., and CLEMENT, J., eds. (Franklin Institute Press, Philadelphia).
- CLEMENT, J. (1982), "Students' Preconceptions in Introductory Mechanics," *Am. J. Phys.* **50**, 66.
- CLEMENT, J. (1983), "A Conceptual Model Discussed by Galileo and Used Intuitively by Physics Students," in *Mental Models*, GENTNER, D., and STEVENS, A. L., eds. (Lawrence Erlbaum Associates, Hillsdale, NJ).
- CLEMENT, J. (1987), "Overcoming Students' Misconceptions in Physics: The Role of Anchoring Intuitions and Analogical Validity," in *Proceedings of Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics III*, NOVAK, J., ed. (Cornell University, Ithaca, NY).
- CLEMENT, J., LOCHHEAD, J., and MONK, G. S. (1981), "Translation Difficulties in Learning Mathematics," *Am. Mathematical Monthly*, **88**, 286.
- COHEN, I. B., ed. (1941), *B. Franklin's Experiments, a New Edition of Franklin's Experiments and Observations on Electricity*, edited, with a critical and historical introduction (Harvard University Press, Cambridge, MA). [Letter IV from this edition is reprinted in *Science* **123**, 47 (1956) in recognition of the 250th anniversary of the birth of Benjamin Franklin.]
- COHEN, R., EYLON, B., and GANIEL, U. (1983), "Potential Difference and Current in Simple Electric Circuits: A Study of Students' Concepts," *Am. J. Phys.* **51**, 407.
- CONANT, J. B., and NASH, L. K., eds. (1957), *Harvard Case Histories in Experimental Science* (Harvard University Press, Cambridge, MA).
- CRANE, H. R. (1960), "Creative Thinking and Experimenting," *Am. J. Phys.* **28**, 437.
- CRANE, H. R. (1969a), "Better Teaching with Better Problems and Exams," *Phys. Today* **22**(3).
- CRANE, H. R. (1969b), "Problems for Introductory Physics. Part I," *The Physics Teacher* **7**, 371.
- CRANE, H. R. (1970), "Problems for Introductory Physics. Part II," *The Physics Teacher* **8**, 182.
- DI SESSA, A. (1982), "Unlearning Aristotelian Physics: A Study of Knowledge-Based Learning," *Cog. Sci.* **6**, 37.
- DRIVER, R., and EASELEY, J. (1978), "Pupils and Paradigms: A Review of Literature Related to Concepts Development in Adolescent Science Students," *Studies in Science Education* **5**, 61.
- EINSTEIN, A. (1905a), "Über eine die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," *Ann. d. Phys.* **4**, 17, 132. [English translation: Arons and Peppard (1965).]
- EINSTEIN, A. (1905b), "Zur Elektrodynamik bewegter Körper" *Ann. d. Phys.* **4**, 17, 891. [English translation in *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*. (Methuen

- & Co. Ltd., 1923. Reprinted by Dover Publications, Inc., New York.)]
- EINSTEIN, A. (1961), *Relativity: The Special and the General Theory* (Crown Publishers, New York).
- EISBERG, R. M. (1976), *Applied Mathematical Physics with Programmable Pocket Calculators* (McGraw-Hill, New York).
- EISENBUD, L. (1958), "On the Classical Laws of Motion," *Am. J. Phys.* **26**, 144.
- Elementary Science Study (ESS)* (1968ff), (Webster Division, McGraw-Hill Book Co., New York). Available from Delta Education, Nashua, NH. (Information about these materials, together with other elementary science curricula, are available on a CD-ROM published under the title *Science Helper K-8 CD ROM* by PC-SIG, Sunnyvale, CA.)
- ERICKSON, G., and AGUIRRE, J. (1984), "Student Conceptions About the Vector Characteristics of Three Physics Concepts," *J. Res. in Sci. Teach.* **21**(5).
- ERLICHSON, H. (1977), "Work and Kinetic Energy for an Automobile Coming to a Stop," *Am. J. Phys.* **45**, 769.
- EVANS, J. (1978), "Teaching Electricity with Batteries and Bulbs," *The Physics Teacher* **16**, 15.
- FARADAY, M. (1839), "Identity of Electricities Derived from Different Sources," in *Experimental Researches in Electricity, Vol. I* (Taylor and Francis, London). (Reprinted by Dover Publications, New York. 1965a, p. 76.)
- FARADAY, M. (1855), "On the Physical Character of the Lines of Magnetic Force," in *Experimental Researches in Electricity, Vol. III* (Taylor and Francis, London). (Reprinted by Dover Publications, New York. 1965b, p. 408.)
- FEYNMAN, R. P., LEIGHTON, R. B., and SANDS, M. (1963), *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA).
- FERGUSON-HESSLER, M. G. M., and DE JONG, T. (1987), "On the Quality of Knowledge in the Field of Electricity and Magnetism," *Am. J. Phys.* **55**, 492.
- FLANSBURG, E. (1972) "Teaching Objectives for a Liberal Arts Physics Laboratory," *Am. J. Phys.* **40**, 1607.
- FREDETTE, M. H., and CLEMENT, J. J. (1981), "Student Misconceptions of an Electric Circuit: What Do They Mean?" *J. Coll. Sci. Teach.* **10**, 280.
- FRENCH, A. P. (1995), "On Weightlessness," *Am. J. Phys.* **63**(2), 105.
- GALLOWAY III, L. A., and WILSON, J. F. (1992), "Measuring the Mechanical Equivalent of Heat Electrically," *The Physics Teacher* **30**, 504.
- GENTNER, D., and GENTNER, D. R. (1983), "Flowing Water or Teeming Crowds: Mental Models of Electricity," in *Mental Models*, GENTNER, D., and STEVENS, A. L., eds. (Lawrence Erlbaum Associates, Hillsdale, NJ).
- GOLDBERG, F. M., and ANDERSON, J. H. (1989), "Student Difficulties with Graphical Representations of Negative Values of Velocity," *The Physics Teacher* **27**, 254.
- GOLDBERG, F., and BENDALL, S. (1995), "Making the Invisible Visible: A Teaching/Learning Environment that Builds on a New View of the Physics Learner," *Am. J. Phys.* **63**(11), 978-991.
- GOLDBERG, F. M., and McDERMOTT, L. C. (1986), "Student Difficulties in Understanding Image Formation by a Plane Mirror," *The Physics Teacher* **24**, 472.
- GOLDBERG, F. M., and McDERMOTT, L. C. (1987), "An Investigation of Student Understanding of the Real Image Formed by a Converging Lens or Concave Mirror," *Am. J. Phys.* **55**, 108.

- GOLDSTEIN, M., and GOLDSTEIN, I. F. (1978), *How We Know* (Plenum Press, New York).
- GUNSTONE, R. F. (1984), "Circular Motion: Some Pre-Instructional Alternative Frameworks," *Res. in Sci. Ed.* **14**, 125.
- GUNSTONE, R. F. (1987), "Student Understanding in Mechanics: A Large Population Survey," *Am. J. Phys.* **55**, 691.
- GUNSTONE, R. F., CHAMPAGNE, A. B., and KLOPFER, L. E. (1981), "Instruction for Understanding: A Case Study," *Austral. Sci. Teachers Jour.* **27**, 27.
- GUNSTONE, R. F., and WHITE, R. (1981), "Understanding Gravity," *Sci. Ed.* **65**, 291.
- HABER-SCHAIM, U., et al. (1960ff), *PSSC Physics*, Several Editions 1960 to the present (D.C. Heath & Co., Lexington, MA).
- HAKE, R. R. (1987), "Promoting Student Crossover to the Newtonian World," *Am. J. Phys.* **55**, 878.
- HAKE, R. R. (1992), "Socratic Pedagogy in the Introductory Physics Laboratory," *The Physics Teacher.* **30**, 546.
- HALLOUN, I. A., and HESTENES, D. (1985), "The Initial State of College Physics Students," *Am. J. Phys.* **53**, 1043; also "Common Sense Concepts About Motion," *Am. J. Phys.* **53**, 1056.
- HALLOUN, I. A., and HESTENES, D. (1987), "Modeling Instruction in Mechanics," *Am. J. Phys.* **55**, 455.
- HARNWELL, G. P., and LIVINGOOD, J. J. (1933), *Experimental Atomic Physics* (McGraw-Hill, New York).
- HEILBRON, J. L. (1979), *Electricity in the 17th and 18th Centuries. A Study of Early Modern Physics* (University of California Press, Berkeley, CA).
- HELLER, J. I., and REIF, F. (1984), "Prescribing Effective Human Problem Solving Processes: Problem Description in Physics," *Cog. and Instr.* **1**, 177.
- HELLER, P., KEITH, R., and ANDERSON, S. (1992), "Teaching Problem Solving Through Cooperative Grouping. Part 1: Group Versus Individual Problem Solving," *Am. J. Phys.* **60**(7), 627-36.
- HELLER, P., and HOLLABAUGH, M. (1992), "Teaching Problem Solving Through Cooperative Grouping. Part 2: Designing Problems and Structuring Groups," *Am. J. Phys.* **60**(7), 637-44.
- HESTENES, D. (1987), "Toward a Modeling Theory of Physics Instruction," *Am. J. Phys.* **55**, 440.
- HOBBIE, R. K. (1973), "Teaching Exponential Growth and Decay: Examples from Medicine," *Am. J. Phys.* **41**, 389.
- HOFSTADTER, D. R. (1982), "Number Numbness or Why Innumeracy May Be Just as Dangerous as Illiteracy," *Sci. Am.* **5**, 20.
- HOLTON, G. (1973), *Introduction to Concepts and Theories in Physical Science*, 2nd ed. revised by BRUSH, S. G. (Addison-Wesley, Reading, MA).
- HOLTON, G. (1978), "Subelectrons, Presuppositions, and the Millikan-Ehrenhaft Dispute," in *The Scientific Imagination* (Cambridge University Press, Cambridge).
- HOOPER, W. (1993), Letter "No Cal," *The Physics Teacher.* **31**, 68.
- HUGGINS, E. R. (1968), *Physics 1* (W. A. Benjamin, New York).
- HUGHES, A. L. (1913), "On the Emission Velocities of Photo-Electrons," *Phil. Trans. Roy. Soc. London, Series A* **212**, 205.

- HUMPHREY, R. D. (1951), *Georges Sorel, Prophet Without Honor* (Harvard University Press, Cambridge, MA).
- INHELDER, B., and PIAGET, J. (1958), *The Growth of Logical Thinking from Childhood to Adolescence* (Basic Books, New York).
- IONA, M. (1979), "Teaching Electrical Resistance," *The Physics Teacher* **17**, 299.
- IONA, M. (1983), "We Ought to Use the Conventional Current Direction," *The Physics Teacher* **21**, 334.
- IONA, M. (1987), "Why Johnny Can't Learn Physics from Textbooks I Have Known," *Am. J. Phys.* **55**, 299.
- IVANY, J. W. G., and PARLETT, M. R. (1968), "The Divergent Laboratory," *Am. J. Phys.* **36**, 1072.
- JOHNSTONE, A. H., and MUGHOL, A. R. (1978), "The Concept of Electrical Resistance," *Phys. Educ.* **13**, 46.
- KARPLUS, R. (1977), "Science Teaching and the Development of Reasoning," *J. Res. Sci. Ed.* **14**, 169.
- KARPLUS, R., KARPLUS, E., FORMISANO, M., and PAULSEN, A-C. (1979), "Proportional Reasoning and Control of Variables in Seven Countries," in *Cognitive Process Instruction*, LOCHHEAD, J., and CLEMENT, J., eds. (Franklin Institute Press, Philadelphia).
- KELLER, J. B. (1987), "Newton's Second Law," *Am. J. Phys.* **55**, 1145.
- KENNEDY, G. (1955), *Education at Amherst* (Harper and Brothers, New York).
- KIDD, R., ARDINI, J., and ANTON, A. (1989), "Evolution of the Modern Photon," *Am. J. Phys.* **57**, 27.
- KING, J. G. (1966), "On Physics Project Laboratories," *Am. J. Phys.* **34**, 1058.
- KRUGLAK, H. (1952) (a) "Experimental Outcomes of Laboratory Instruction in Elementary College Physics," *Am. J. Phys.* **20**, 136; (b) "Achievement of Physics Students With and Without Laboratory Work," *Am. J. Phys.* **21**, 14.
- KUEHL, E. (1984), (Chairman The Evaluation Committee), "An Evaluation of High School Physics Laboratory Manuals," *The Physics Teacher* **22**, 222.
- KUNZ, K. S. (1971), "Visualizing Large Numbers," *Am. J. Phys.* **39**, 452.
- LAMB, H. (1932), *Hydrodynamics* (Cambridge University Press, Cambridge).
- LAPP, C. J. (1940), "Effectiveness of Mathematical versus Physical Solutions in Problem Solving in College Physics," *Am. J. Phys.* **8**, 241.
- LARKIN, J. (1981), "Cognition of Learning Physics," *Am. J. Phys.* **49**, 534.
- LARKIN, J. (1983), "The Role of Problem Representation in Physics," in *Mental Models*, GENTNER, D., and STEVENS, A. L., eds. (Lawrence Erlbaum Associates, Hillsdale, NJ).
- LAWS, P. (1991a), "Calculus-Based Physics Without Lectures," *Physics Today* **44** (12), 24-31.
- LAWS, P. (1991b), "Workshop Physics: Learning Introductory Physics by Doing It," *Change* July/August.
- LAWSON, A. E. (1980), "Relationships among Level of Intellectual Development, Cognitive Style, and Grades in a College Biology Course," *Sci. Ed.* **64**, 95.
- LAWSON, A. E. (1982), "The Reality of General Cognitive Operations," *Sci. Ed.* **66**, 229.
- LAWSON, A. E., LAWSON, D. I., and LAWSON, C. A. (1984), "Proportional Reasoning and the Linguistic Abilities Required for Hypothetico-Deductive Reasoning," *J. Res. Sci. Teach.* **21**, 119.

- LAWSON, A. E., and WOLLMAN, W. T. (1975), "Physics Problems and the Process of Self-Regulation," *The Physics Teacher* **13**, 470.
- LAWSON, R. A., and McDERMOTT, L. C. (1987), "Student Understanding of the Work-Energy and Impulse-Momentum Theorems," *Am. J. Phys.* **55**, 811.
- LENARD, P. (1902), "Über die Lichtelektrische Wirkung," *Ann. d. Phys.* **4**, 8, 149.
- LIN, H. (1978), "Newtonian Mechanics and the Human Body: Some Estimates of Performance," *Am. J. Phys.* **46**, 15.
- LIN, H. (1982), "Fundamentals of Zoological Scaling," *Am. J. Phys.* **50**, 72.
- LOCHHEAD, J. (1981), "Faculty Interpretations of Simple Algebraic Statements: The Professor's Side of the Equation," *Jour. of Mathematical Behavior* Spring 1981.
- LONG, D. D., McLAUGHLIN, G. W., and BLOOM, A. M. (1986), "The Influence of Physics Laboratories on Student Performance in a Lecture Course," *Am. J. Phys.* **54**, 122.
- MACH, E. (1893), *The Science of Mechanics* (Open Court Publishing Co., Chicago).
- MAGIE, W. F. (1935), *Source Book in Physics* (McGraw-Hill, New York).
- MALONEY, D. P. (1984), "Rule Governed Approaches to Physics: Newton's Third Law," *Phys. Ed.* **19**, 37.
- McCLOSKEY, M., CAMARAZZA, A., and GREEN, B. (1980), "Curvilinear Motion in the Absence of External Forces," *Science* **210**, 1139.
- McCLOSKEY, M. (1983), "Intuitive Physics," *Sci. Am.* **249**, 122.
- McDERMOTT, L. C. (1980), "Teaching Physics to Promote Cognitive Development in Academically Disadvantaged Students Aspiring to Science-Related Careers," in *Physics Teaching*. Proceedings of the GIREP conference held at the Weizmann Institute of Science, Rehovot, Israel. August 19-24, 1979. GANIEL, U., ed. (Balaban International Science Services, Jerusalem).
- McDERMOTT, L. C. (1984), "Research on Conceptual Understanding in Mechanics," *Physics Today* **37**, 24.
- McDERMOTT, L. C. (1990), "A Perspective on Teacher Preparation in Physics and Other Sciences: The Need for Special Courses for Teachers," *Am. J. Phys.* **58**(5), 452.
- McDERMOTT, L. C. (1991), "What We Teach and What is Learned: Closing the Gap," *Am. J. Phys.* **59**(4), 301.
- McDERMOTT, L. C. (1993), "Guest Comment: How We Teach and How Students Learn—A Mismatch?" *Am. J. Phys.* **61**(4), 295.
- McDERMOTT, L. C., PITERNICK, L. K., and ROSENQUIST, M. L. (1980), "Helping Minority Students Succeed in Science. I. Development of a Curriculum in Physics and Biology," *J. Coll. Sci. Teach.* **9**, 135.
- McDERMOTT, L. C., ROSENQUIST, M. L., and VAN ZEE, E. H. (1983), "Strategies to Improve the Performance of Minority Students in the Sciences," in *Teaching Minority Students: New Directions for Teaching and Learning*, No. 16. CONES, J. H. III, NOONAN, J. F., and JANHA, D., eds. (Jossey-Bass, San Francisco).
- McDERMOTT, L. C., ROSENQUIST, M. L., and VAN ZEE, E. H. (1987), "Student Difficulties in Connecting Graphs and Physics: Examples from Kinematics," *Am. J. Phys.* **55**, 503.
- McDERMOTT, L. C., and SHAFFER, P. (1992), (a) "Research as a Guide for Curriculum Development: An Example from Introductory Electricity, Part I: Investigation of Student Understanding," *Am. J. Phys.* **60**(11), 994 [see Erra-

- tum to Part I, *Am. J. Phys.* **61**(1), 81 (1993)); (b) "Part II: Design of Instructional Strategies," *Am. J. Phys.* **60**, 994-1013.
- McDERMOTT, L. C., SHAFFER, P., and SOMMERS, M. (1994), "Research as a Guide for Curriculum Development: An Illustration in the Context of the Atwood's Machine," *Am. J. Phys.* **62**(1), 46.
- McDERMOTT, L. C., SHAFFER, P., and ROSENQUIST, M. (1996), *Physics by Inquiry* (John Wiley & Sons, Inc. New York).
- McKINNON, J. W., and RENNER, J. W. (1971), "Are Colleges Concerned with Intellectual Development?" *Am. J. Phys.* **39**, 1047.
- MEMORY, J. D. (1973), "Kinematics Problems for Joggers," *Am. J. Phys.* **41**, 1205.
- MEMORY, J. D., and JENKINS, A. W., Jr. (1977), "Estimating Orders of Magnitude," *The Physics Teacher* **15**, 43.
- MERRILL, J. R. (1976), *Using Computers in Physics* (Houghton Mifflin Co., Boston).
- MICHELS, W. C. (1962), "Role of Experimental Work," *Am. J. Phys.* **30**, 172.
- MILLIKAN, R. A. (1909), "A New Modification of the Cloud Method of Measuring the Elementary Electrical Charge and the Most Probable Value of that Charge," *Phys. Rev.* **29**, 560.
- MILLIKAN, R. A. (1911), "The Isolation of an Ion, a Precision Measurement of Its Charge, and the Correction of Stokes's Law," *Phys. Rev.* **32**, 349.
- MILLIKAN, R. A. (1916), "A Direct Photoelectric Determination of Planck's Constant 'h'," *Phys. Rev.* **7**, 355.
- MILLIKAN, R. A. (1917), *The Electron* (University of Chicago Press, Chicago (1st ed. 1917; 2nd ed. 1924)).
- MILLIKAN, R. A. (1949), "Albert Einstein on his Seventieth Birthday," *Rev. Mod. Phys.* **21**, 343.
- MINSTRELL, J. (1982), "Explaining the 'At Rest' Condition of an Object," *The Physics Teacher* **20**, 10.
- MINSTRELL, J. (1984), "Teaching for the Understanding of Ideas: Forces on Moving Objects," in *1984 Yearbook of the Association for the Education of Teachers* (ERIC Clearinghouse, Ohio State University, Columbus, OH).
- MOREIRA, M. A. (1980), "A Non-Traditional Approach to Evaluation of Laboratory Instruction in General Physics Courses," *Int. J. Sci. Ed.* **2**, 441.
- MORRISON, P. (1963), "Fermi Questions," *Am. J. Phys.* **31**, 626.
- OKUN, L. B. (1989), "The Concept of Mass," *Physics Today* **42**(6), 31.
- OPPENHEIMER, J. R. (1963), "Communication and Comprehension of Scientific Knowledge," *Science* **142**, 1144.
- OSBORNE, R. (1984), "Children's Dynamics," *The Physics Teacher* **22**, 504.
- PANOFSKY, W. K. H., and PHILLIPS, M. (1962), *Classical Electricity and Magnetism*, 2nd ed. (Addison-Wesley, Reading, MA).
- PENCHINA, C. M. (1978), "Pseudowork-Energy Principle," *Am. J. Phys.* **46**, 295.
- PERRY, W. G. (1970), *Forms of Intellectual and Ethical Development in the College Years* (Holt, Rinehart & Winston, New York).
- PETERS, P. C. (1982), "Even Honors Students Have Conceptual Difficulties With Physics," *Am. J. Phys.* **50**, 501.
- PETERS, P. C. (1986), "An Alternative Derivation of Relativistic Momentum," *Am. J. Phys.* **54**, 804.
- PFISTER, H., and LAWS, P. (1995), "Kinesthesia-1: Apparatus to Experience 1-D Motion," *The Physics Teacher* **33**(4), 214-220.

- PHILLIPS, M. (1981) "Early History of Physics Laboratories for Students at the College Level," *49*, 522.
- PIAGET, J., and INHELDER, B. (1958), *Growth of Logical Thinking* (Basic Books, New York).
- PRESCOTT, J. R., and ANGER, C. D. (1970) "Removing the 'Cook Book' From Freshman Physics Laboratories," *Am. J. Phys.* **38**, 58.
- PSSC Physics films (ca. 1958-59), currently available under title "Physics Cinema Classics" from Ztek Co., P.O. Box 11768, Lexington, KY 40577.
- REID, W. M., and ARSENEAU, D. F. (1971), "Labs of Unlimited Scope," *Am. J. Phys.* **39**, 271.
- REIF, F. (1981), "Teaching Problem Solving—A Scientific Approach," *The Physics Teacher* **19**, 310.
- REIF, F. (1985), "Acquiring an Effective Understanding of Scientific Concepts," in *Cognitive Structure and Conceptual Change*, WEST, L. H. T., and PINES, L., eds. (Academic Press, Orlando, FL).
- REIF, F. (1995), "Understanding and Teaching Important Scientific Thought Processes," *Am. J. Phys.* **63** (1), 17.
- REIF, F., and HELLER, J. I. (1982), "Knowledge Structures and Problem Solving in Physics," *Educational Psychologist* **17**, 102.
- REIF, F., and HELLER, J. I. (1984), "Prescribing Effective Human Problem Solving Processes: Problem Description in Physics," *Cognition and Instruction* **1**, 177.
- REIF, F., LARKIN, J. H., and BRACKETT, B. C. (1976), "Teaching General Learning and Problem Solving Skills," *Am. J. Phys.* **44**, 212.
- REIF, F., and ST. JOHN, M. (1979) "Teaching Physicists' Thinking Skills in the Laboratory," *Am. J. Phys.* **47**, 950.
- RESNICK, R., and HALLIDAY, D. (1977, 1985), *Physics*, 3rd and 4th eds. (John Wiley & Sons, New York).
- RESNICK, R., HALLIDAY, D., and KRANE, K. S. (1992) *Physics*. (John Wiley & Sons, Inc., New York).
- RICHARDSON, O. W., and COMPTON, K. T. (1912), "The Photoelectric Effect," *Phil. Mag. Series 6*, **24**, 575.
- ROBINSON, M. C. (1979), "Undergraduate Laboratories in Physics: Two Philosophies," *Am. J. Phys.* **47**, 859.
- ROGERS, E. M. (1960), *Physics for the Inquiring Mind* (Princeton University Press, Princeton, NJ).
- ROLLER, D., and ROLLER, D. H. D. (1957), "The Development of the Concept of Electrical Charge," in *Harvard Case Histories in Experimental Science*, CONANT, J. B., and NASH, L. K., eds. (Harvard University Press, Cambridge, MA).
- ROMER, R. H. (1982), "What Do 'Voltmeters' Measure?: Faraday's Law in a Multiply Connected Region," *Am. J. Phys.* **50**, 1089.
- ROSENQUIST, M. L. (1982), *Improving Preparation for College Physics of Minority Students Aspiring to Science-Related Careers* (Unpublished dissertation, University of Washington, Seattle, WA).
- ROSENQUIST, M. L., and McDERMOTT, L. C. (1987), "A Conceptual Approach to Teaching Kinematics," *Am. J. Phys.* **55**, 407.
- ROSNICK, P., and CLEMENT, J. (1980), "Learning Without Understanding: The Effect of Tutoring Strategies on Algebra Misconceptions," *Jour. of Mathematical Behavior* **3**, No. 1.

- RUTHERFORD, E. (1903), "The Magnetic and Electric Deviation of the Easily Absorbed Rays from Radium," *Phil. Mag.* 5, 6, 177.
- RUTHERFORD, E. (1906), "The Mass and Velocity of the α -Particles Expelled from Radium and Actinium," *Phil. Mag.* 12, 6, 348.
- RUTHERFORD, E., and ROYDS, T. (1909), "The Nature of the α -Particle from Radioactive Substances," *Phil. Mag.* 17, 6, 281 [reprinted in Birks (1962)].
- RUTHERFORD, F. J., HOLTON, G., and WATSON, F. G. (1981), *The Project Physics Course*, 3rd ed. (Holt, Rinehart and Winston, New York).
- Science Curriculum Improvement Study (SCIS)* (1968ff, Delta Education, Nashua, NH.) [Information about these materials, together with other elementary science curricula, are available on a CD-ROM published under the title *Science Helper K-8 CDROM* by PC-SIG, Sunnyvale, CA.]
- SHAHN, E. (1988), "On Science Literacy," in *Educational Philosophy and Theory*. Journal of the Philosophy of Education Society of Australia. Special topic issue on Science Education. MATTHEWS, M. R., ed. (20)2, 42.
- SHAMOS, M. (1959), *Great Experiments in Physics* (Henry Holt & Co., New York).
- SHANKLAND, R. S. (1963), "Conversations with Albert Einstein," *Am. J. Phys.* 31, 47.
- SHERWOOD, B. A. (1983), "Pseudowork and Real Work," *Am. J. Phys.* 51, 597.
- SHERWOOD, B. A., and BERNARD, W. H. (1984), "Work and Heat Transfer in the Presence of Sliding Friction," *Am. J. Phys.* 52, 1001.
- SHONLE, J. I. (1970), "A Progress Report on Open-End Laboratories," *Am. J. Phys.* 38, 450.
- SHYMANSKY, J. A., KYLE, W. C., JR., and ALPORT, J. M. (1983), "The Effects of New Science Curricula on Student Performance," *J. Res. Sci. Teach.* 20, 387.
- SPEARS, J., and ZOLLMAN, D. (1977), "The Influence of Structured Versus Unstructured Laboratory on Students' Understanding of the Process of Science," *J. Res. Sci. Teach.* 14, 33.
- ST. JOHN, M. (1980), "Thinking Like a Physicist in the Laboratory," *The Physics Teacher* 18, 436.
- STEAD, B. F., and OSBORNE, R. J. (1980), "Exploring Science Students' Concepts of Light," *Austr. Sci. Teach. Jour.* 26, 84.
- STEINBERG, M. S. (1983), "Reinventing Electricity," in *Proceedings of the International Seminar on Misconceptions in Science and Mathematics*, HELM, H. and NOVAK, J., eds. (Cornell University, Ithaca, NY).
- STEINBERG, M. S. (1987), "Transient Lamp Lighting with High-Tech Capacitors," *The Physics Teacher* 25, 95.
- STEVENSON, H. W., SHIN-YING LEE, and STIGLER, J. W. (1986), "Mathematics Achievement of Chinese, Japanese, and American Children," *Science* 231, 693.
- STRIKE, K. A., and POSNER, G. J. (1982), "Conceptual Change and Science Teaching," *Eur. J. Sci. Ed.* 4, 231.
- SWARTZ, C. E., and ZIPFEL, C. (1972), "Individualized Instruction in Introductory Physics," *Am. J. Phys.* 40, 1436.
- THOMSON, G. P. (1956), "J. J. Thomson and the Discovery of the Electron," *Physics Today* 9(8), 19.
- THOMSON, J. J. (1897), "Cathode Rays," *Phil. Mag.* 44, 5, 293 [Excerpts are to be found in Magie (1935) and Shamos (1959)].

- THOMSON, J. J. (1899), "On the Masses of Ions in Gases at Low Pressures," *Phil. Mag.* **48**, 5, 547.
- THOMSON, J. J. (1912), "Multiply Charged Atoms," *Phil. Mag.* **24**, 6, 668.
- THORNTON, R. K. (1987), "Tools for Scientific Thinking: Microcomputer-Based Laboratories for Physics Teaching," *Phys. Ed.* **22**, 230.
- THORNTON, R. K., and SOKOLOFF, D. A. (1990), "Learning Motion Concepts Using Real-Time Microcomputer-Based Laboratory Tools," *Am. J. Phys.* **58**, 858-867.
- TIPLER, P. A. (1976, 1982), *Physics* (Worth Publishers, New York).
- TOBIAS, S. (1986), "Peer Perspectives on the Teaching of Science," *Change* **18**, 36.
- TOBIAS, S., and HAKE, R. R. (1988), "Professors as Physics Students: What Can They Teach Us?" *Am. J. Phys.* **56**, 786.
- TOLMAN, R. C., and STEWART, T. D. (1916), "The Electromotive Force Produced by the Acceleration of Metals," *Phys. Rev.* **8**, 97.
- TOLMAN, R. C., and STEWART, T. D. (1917), "The Mass of the Electric Carrier in Copper, Silver, and Aluminum," *Phys. Rev.* **9**, 164.
- TOOTHACKER, W. S. (1983), "A Critical Look at Introductory Laboratory Instruction," *Am. J. Phys.* **51**, 516.
- TROWBRIDGE, D. E., "Graphs & Tracks," Available from Physics Academic Software, Box 8202, North Carolina State University, Raleigh, NC 27695-8202.
- TROWBRIDGE, D. E. (1988), "Applying Research Results to the Development of Computer Assisted Instruction," in *Proceedings of the Conference on Computers in Physics Instruction*, RISLEY, J. S., and REDDISH, E. F., eds. (North Carolina State University, Raleigh, NC).
- TROWBRIDGE, D. E., and McDERMOTT, L. C. (1980), "Investigation of Student Understanding of the Concept of Velocity in One Dimension," *Am. J. Phys.* **48**, 1020.
- TROWBRIDGE, D. E., and McDERMOTT, L. C. (1981), "Investigation of Student Understanding of the Concept of Acceleration in One Dimension," *Am. J. Phys.* **49**, 242.
- TSANTES, E. (1974), "Note on the Tides," *Am. J. Phys.* **42**, 330.
- VIENNOT, L. (1979), "Le Raisonnement Spontane en Dynamique Elementaire," *Eur. J. Sci. Ed.* **1**, 205.
- WATTS, D. M. (1985), "Student Conceptions of Light: A Case Study," *Phys. Ed.* **20**, 183.
- WEBER, F. N. (1992), "Measuring the Mechanical Equivalent of Heat—Mechanically," *The Physics Teacher* **30**, 507.
- WEINSTOCK, R. (1961), "What's F ? What's m ? What's a ?" *Am. J. Phys.* **29**, 698.
- WEINSTOCK, H., and BORK, A., eds. (1986), *Design of Computer-Based Learning Materials*, Proceedings of the NATO Advanced Study Institute on Learning Physics and Mathematics via Computers held in San Miniato, Italy, 15-26 July 1985. NATO. ASI Series F, *Computer and Systems Sciences*, Vol. 23 (Springer Verlag, Berlin, New York).
- WELLMAN, R. T. (1978), "Science: A Basic Language for Reading Development," in *What Research Says to the Science Teacher* Vol. 1., ROWE, M. B., ed. (National Science Teachers Association, Washington, DC).
- WESTFALL, R. S. (1971), *Force in Newton's Physics* (American Elsevier, New York).

- WHITAKER, R. J. (1983), "Aristotle Is not Dead: Student Understanding of Trajectory Motion," *Am. J. Phys.* **51**, 352.
- WHITE, B. (1983), "Sources of Difficulty in Understanding Newtonian Dynamics," *Cog. Sci.* **7**, 41.
- WHITE, B. (1984), "Designing Computer Games to Help Physics Students Understand Newtonian Laws of Motion" *Cog. and Instr.* **1**, 69.
- WHITE, R. (1979), "Relevance of Practical Work to Comprehension of Physics," *Phys. Ed.* **14**, 384.
- WHITEHEAD, A. N. (1929), *The Aims of Education* (Macmillan, New York).
- WHITTAKER, E. (1951), *A History of the Theories of Aether and Electricity* (Thos. Nelson & Sons, London; also Harper Torchbook No. 531, Harper & Brothers, New York, 1960).
- WILLIAMS, E. R., FALLER, J. E., and HILL, H. A. (1971), "New Experimental Test of Coulomb's Law: A Laboratory Upper Limit on the Photon Rest Mass," *Phys. Rev. Lett.* **26**, 721.
- WISER, M., and CAREY, S. (1983), "When Heat and Temperature Were One," in *Mental Models*, GENTNER, D., and STEVENS, A. L., eds. (Lawrence Erlbaum Associates, Hillsdale, NJ).

Index

A

- Acceleration, 32ff
 - continuity equation for, 99
 - Galileo's choice of $\Delta v/\Delta t$, 35
 - misleading treatment of, 24
 - research on concept of, 42ff
- Action at a distance, 79
 - and electrostatic interaction, 183
 - and lines of force, 228ff
 - Faraday and, 79, 227ff, 356
 - Maxwell and, 80, 231ff
 - field theory and, 80
- Algebra
 - and ratio reasoning, 8
 - interpreting algebraic statements, 17
- Alpha particles
 - identified as helium, 281
 - Rutherford scattering of, 284
- Algebraic signs, 32
 - of Δv , 34
- Ampere's experiment, 223ff
 - evidence against electrostatic explanation, 224
- Area, 1ff
 - in kinematics graphs, 36
 - operational definition, 2
 - relation to "integral", 3, 36
- Arithmetical reasoning, 8ff
 - and division, 8
 - and graphs, 9
- Arrows
 - distinguishing F , v , and a , 81
- Atoms
 - nuclei, 284
 - size of, 282
 - spacing of, 283
- Avogadro's number, 268ff

B

- "Backwards" science, 178ff, 197, 289

- Balmer formula, 280

"Because"

- misleading usage, 82, 177ff

B-field

- measurement of, 233

Bohr atom, 292ff

- written homework on, 313ff

C

- Center of mass
 - revolution around, 128ff
- Centripetal force, 121ff, 161ff
 - Newton's derivation, 161
 - with colinear forces, 121
 - with non-colinear forces, 124ff
- Centrifugal force
 - as a fictitious force, 127
- Charge, electric, 168ff
 - and electrons, 170
 - conservation of, 184
 - experiments at home, 170
 - like and unlike, 171
 - positive and negative, 174ff
 - quantification of, 179ff
- Chaos, 342
- Circuit, electric, 194ff
 - applying the model, 198ff
 - building the model, 206ff
 - experiments at home, 198
 - freedom of charge carriers in, 205, 211
 - two-endedness of components, 194ff
- Circular motion
 - preconceptions regarding, 119
- Clock reading
 - versus "time", 24ff
- Collisions, 157, 284, 324
- Computer
 - use in kinematics, 39
 - Socratic dialogs with, 369
- Correspondence principle, 297ff, 315

Coulomb's law, 179ff
 Critical thinking, 184, 375ff
 list of processes, 376ff

D

Data analysis, 331ff
 accuracy, 331
 average, 331
 precision, 331
 significant differences, 331
 Declarative knowledge, 347ff, 376
 Direct proportion
 versus linear relation, 12
 Discharge tubes
 demonstrations with, 267
 Distance
 versus "position", 24ff
 Distribution function, 332
 Division, 4ff, 330
 counting subtractions, 4

E

Electric field
 normal to conductor surface, 92
 strength, 185ff
 superposition of, 186
 Electricity
 current, 188ff
 applying the model, 198ff
 bulk or surface phenomenon?, 205ff
 conductors and nonconductors, 195
 dynamic nature of, 195
 inventing the model, 194ff, 206ff
 Joule's law, 201ff
 motion of charge in, 189, 190ff
 identity from different sources, 190
 power transmission, 210
 within atoms, 268
 Electromagnetism, 218ff
 and light, 229
 computer-based drill, 225
 mnemonics for, 225
 Electron, 170, 179, 272ff, 277, 312
 charge of, 278ff
 free in metals, 210ff
 in current electricity, 208ff
 Electroscope, 171
 Empirical
 versus theoretical, 280

Energy, 135ff, 146ff
 conservation of, 202
 in everyday phenomena, 137
 in radioactivity, 268
 in rolling down incline, 154ff
 internal, 146
 law of conservation, 146
 under zero-work forces, 153ff
 Enthalpy, 148ff
 Equals signs
 multiple meanings of, 97ff
 Estimating, 15, 283, 285, 329
 Event, 25
 Exponential
 growth and decay, 283

F

Faculty development, 390ff
 Faraday's law
 in multiply connected region, 226ff
 Field theory
 and action at a distance, 80, 227ff, 356
 and electrostatic interaction, 183
 Field strength
 electrical, 185ff
 gravitational, 185
 First law of thermodynamics, 143, 145
 Force, 57ff
 centrifugal, 127
 centripetal, 121ff
 contact versus non-contact, 76
 distributed and concentrated, 80
 fictitious, 117ff
 free body diagrams, 77
 normal, 90ff
 operational definition, 61
 orthogonal components of, 92, 93, 133
 passive versus active, 76, 122
 redefinition of, 354
 verbal description, 78
 Frames of reference, 127ff, 303
 inertial versus non-inertial, 93
 Free fall, 86
 Friction, 94ff, 115, 116
 and normal forces, 95
 as an accelerating force, 95
 as a passive force, 94

G

Galileo

- and choice of $\Delta v/\Delta t$, 35
- and modern science, 46ff

Graphs

- and arithmetical reasoning, 10ff
- and kinesthetic experience, 28ff
- interpretation of, 29
- meaningless aspects, 29, 30
- s versus t , 28
- v versus t , 35

Gravity, 81ff

- and presence of air, 83
- confusion with g , 34
- Galileo's comment on, 82

H

Heat, 139ff

- and temperature, 139
- and thermal internal energy, 148
- converting work into, 142

Heuristic device, 230, 290 356

Horizontal, 16, 83

Hydrostatic pressure, 80

Hypothesis

- testing of, 47, 345, 353

Hypothetico-deductive reasoning, 381

I

- Idea first, name afterwards, 27, 33, 82, 110, 117, 131, 134, 177, 185, 195, 200, 345, 371, 379, 386

Ideal gas

- kinetic model of, 320ff

Images, 258ff

- thin lenses, 260ff
- plane mirror, 258ff

Impulse, 142ff

- and area, 3
- impulse-momentum theorem, 142ff

Induction

- charging by, 179

Inductive reasoning, 181, 347, 381

Inert ideas, 334, 377

Inertia

- law of, 59, 69
- demonstrations of, 96

Instant, 24ff

- "for an" versus "at an", 31
- versus "time", 25ff

Integral

- and area, 3

Interaction

- and Newton's third law, 78
- thermal, 140
- electrical, 167

Interference

- grating versus two-slit, 255

Ion beams, 274

Ionization

- of gases, 267

Irrelevant data

- in problems, 99

J

Joule's law, 201ff

Joule heat, 201ff

K

Kinematics

- rectilinear, 23ff
- rotational, 118ff

Kinetic theory

- assumptions of, 319ff

L

Laboratory work, 333ff

- guidance in, 335ff

Language

- and arithmetical reasoning, 9
- and literacy, 19
- and operational definition, 18
- and verbal reasoning, 19

Lasers

- interference demonstrations with, 255

Law of multiple proportions, 290, 357

Lenses, 260ff

Light

- and electromagnetism, 229
- interference, 255
- novice conception of, 263
- photon concept, 289ff, 300

Linear relation, 9ff

- vs direct proportion, 12

Linguistic problems, 18, 25, 64, 67, 70, 73ff, 136

Lines of force, 227ff, 356

Literacy, scientific 50, 111, 266, 268, 277, 279, 280, 289, 290, 319, 344ff

components of, 345-346

M

Mathematical physics, 339ff

Mass

confusion with volume, 4
gravitational versus inertial, 67ff
inertial, 57ff
misuse of term, 64
operational definition, 63

Mathematics

avoidance of, 21

Mechanical equivalent of heat, 337

Midnight, 16

Misconceptions, 56

Millikan experiment, 278

Mirrors, 258ff

Momentum, 135ff

in everyday phenomena, 137
force and rate of change of, 138, 162

N

Newton's rings, 256ff

Noon, 16

North-south, direction, 16, 176

Not the case, visualizing what is, 62, 68, 93, 94, 96, 97, 109, 117, 122, 132, 173, 205, 206, 246, 274, 288, 322

Nucleus, atomic, 284

O

Observation and inference, 50, 141, 257, 267, 334, 345, 378

Oersted's experiment, 219ff

Ohm's law, 198, 200ff

and Joule's law, 201ff
teaching of, 204ff

Operational definition, 27, 58, 379

and language, 18

of area, 2

of electric, magnetic, and gravitational effects, 168ff, 176

of horizontal, 16

of like and unlike charges, 172ff

of magnetic poles, 174ff

of midnight, 16

of noon, 16

of transfer of l. 141

of vertical, 16

of volume, 3

Operations (Piagetian terminology), 7

Operative knowledge, 347ff, 376

"Over"

duration versus division, 32ff

P

Pascal's law, 238

Per, 7, 32

meaning of, 7

Phenomenological reasoning, 114ff

Photoelectric effect, 285ff

Einstein's equation, 292

Lenard's experiments, 286ff

Photon concept, 289ff, 300

Pi (π), 116

Piltdown hoax, 357

Polarization, electrical, 177ff

Poles

magnetic and terrestrial, 174ff

Position, 24, 25ff

instantaneous, 26

versus "distance", 24ff

Potential difference

in electric circuits, 200, 203ff, 209ff

Preconceptions, 56

Pressure, 238

hydrostatic, 327ff

Principle of detailed balancing, 324ff

Problem solving

in dynamics, 99ff

in kinematics, 38

Projectile motion, 111ff

independence of components, 112

superposition in, 111

vectors in, 113

Protoconcepts, 40ff

Pseudowork, 145ff

Q

Quantum. 279, 285, 292, 294, 295, 317, 326

R

Radian measure, 15, 116ff

units and dimensions of, 117

Radioactivity, 267ff, 280

and energy conservation, 268

and mass conservation, 268

Ratio reasoning
 and electric current, 206
 and scaling, 12, 322, 329
 and trigonometry, 15
 interpreting ratios, 4ff
 with lenses, 264
 Redefining concepts, 60, 67, 141, 150,
 189, 205, 354, 379
 Relativity, 301ff
 length contraction, 306
 simultaneity, 301ff, 305
 Resistance, electrical
 Cavendish's measurements of, 197,
 200, 201
 combinations of resistors, 199ff
 forming the concept, 197
 Joule's law and, 201ff
 teaching of, 204ff
 Reversing line of reasoning, 7, 18, 72,
 241, 251, 254, 262
 "Rigid" objects
 visualizing deformation of, 75
 Rutherford scattering, 284

S

Scaling, 12ff, 283, 322, 329
 and functional relations, 12
 volume, 4
 Second law of motion, 60ff
 Second law of thermodynamics, 157
 Sense perception
 effects eluding direct, 75, 227ff, 321,
 323, 329, 356, 357
 Significant figures, 330ff
 Slope
 as a property, 10
 Slowness, 27
 Sonic rangefinder, 29
 Spiralling back, 10, 57, 65, 70, 78, 80,
 93, 97, 115, 118, 119, 121, 122,
 124, 125, 129, 133, 134, 135,
 138, 161, 167, 168, 182, 184,
 211, 216, 224, 242, 251, 252,
 275, 278, 284, 285, 288, 293,
 300, 303, 328, 385
 Strings
 forces on, 88ff
 massless, 89ff
 Superposition
 Galileo and, 49

 in projectile motion, 111
 of electrical fields, 186
 of masses and forces, 64

T

Teacher education, 365ff, 372
 Tension, 88ff
 operational definition, 89
 Temperature, 139
 distinguishing from heat, 139ff
 thermal equilibrium, 140
 Testing, 390
 Thermal expansion, 329
 Third law of motion, 74ff
 and action at a distance, 79
 and electrostatic interaction, 182
 verbal description of forces, 78
 Thomson's experiment, 272ff
 written homework on, 308ff
 Tides, lunar, 129ff
 Time
 interval, 24
 versus "instant", 25ff
 Tolman-Stewart experiment, 210ff
 Top of flight, 37
 "stopping" at, 37
 Torque, 131ff
 Trigonometry, 15
 Two-body problem, 128ff
 and earth-moon system, 131
 and electron-proton system, 131

U

"Understanding" in science, 113
 Uniform
 meaning of, 31

V

Vacuum, 83
 Variables
 control of, 380
 Vectors, 107ff
 and commutativity, 109
 components of, 110ff
 definition of, 108
 subtraction of, 107
 velocity and acceleration vectors,
 108
 Velocity
 average, 26

- instantaneous, 24, 30ff
- misleading treatment of, 23
- redefinition of, 355
- research on concept of, 40ff

Vertical, 16, 83

Volume

- confusion with mass, 4
- operational definition, 3

W

Waves, 234ff

- fronts, 251ff
- generation of, 237ff
- graphs of, 235ff
- interference, 253ff
 - and Newton's rings, 256ff
 - connecting different physical phenomena, 254
 - grating versus two-source patterns, 2254ff
 - two-source patterns, 253ff
- particle versus propagation velocity, 235
- periodic, 252ff
- rays, 251ff
- reflection of, 238ff, 257
 - diffuse versus specular, 257ff
- sinusoidal, 252ff
- superposition of, 239, 240
- trains, 236, 237
- transient effects, 250ff
- velocity of, 241ff
 - kink on string, 242
 - pulse in fluid, 244
 - wave in shallow water, 247

Weight, 66ff

- and weightlessness, 85
- comparing, 87
- feeling weight of object, 84
- versus mass, 66

"Why" in science, 113

Work, 142ff

- and area, 3
- and pseudowork, 144
- and zero-work forces, 153ff
- work-energy theorem, 142ff

X

X-rays

- ionization caused by, 273

Y

Young's experiment

- with Newton's rings, 256

Z

Zero

- multiple meanings of, 98

PART II

Homework and Test Questions

Preface to Part II

People have now-a-days got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the book from which the lectures are taken.

– Samuel Johnson

‘You damn sadist,’ said mr. cummings, ‘You try to make people think.’

– Ezra Pound, Canto 89

This collection of questions and problems is supplementary to the sample questions and problems found in several sections at the ends of chapters in Part I of this book. The samples in Part I are meant to be illustrative of approaches that lead the student into physical experiences and into sequences of thinking and reasoning that help penetrate learning difficulties. The number of these examples was limited to make Part I slimmer and more readable. Readers should realize that questions in the research protocols that illuminated student preconceptions, misconceptions, and other learning difficulties are themselves very useful when embedded in homework or in tests. To conserve space and limit cost, only a few of the questions and problems suggested in Part I are repeated here, and readers are therefore referred to Part I for these prior materials.

The majority of the questions and problems in this collection have been used and tested either by the author in his own courses or by colleagues who have provided ideas based on their own experience. Some of the questions stem from research experience in our Physics Education Research Group. Many of the questions are especially well suited for use in group discussion and cooperative learning formats.

This collection is confined to a very basic level of subject matter common to the majority of introductory physics courses; more advanced subjects and concepts have been deliberately excluded. Within this range of subject matter, however, the questions range from extremely simple and fundamental to fairly sophisticated. In the latter, fairly extensive guidance is provided. No

attempt is made to duplicate conventional types of numerical end-of-chapter problems readily available in existing texts. The texts are replete with excellent examples. These are a necessary and intrinsic part of our teaching, and this collection is not meant to deprecate or diminish their role. The intent is to fill what researches on teaching and learning show to be gaps in the existing structure. In those instances where fairly conventional questions are being presented, they are generally amplified by addition of Socratic questioning that helps penetration of the question by students who do not otherwise get a start.

These questions and problems are spread over a variety of modes that emerge as essential components in the learning and understanding of physics. Among these modes are forming and applying basic concepts; operational definition; verbalization; connection of abstractions to everyday experience; translating between various representations (e.g., verbal to symbolic or to graphic and the reverse); asking questions; formulating problems; visualizing outcomes in the abstract (hypothetico-deductive reasoning); discriminating between observation and inference; checking for internal consistency; and interpreting results. Students need extensive practice in these modes of thinking and reasoning, and under-supply of such opportunities is one of the principal shortcomings of conventional texts and much physics instruction. This collection is a limited attempt to bridge some of the existing gaps, but I do not pretend that it is either complete or final. Much will be learned through research in the coming years, and much can be added by practicing teachers who bring their own varied imaginations to the building of more comprehensive and more effective collections. No one looks forward more eagerly than I to the additions that invariably emerge when fresh imaginations supplement the limited output of a single individual or a small initial group.

Users of this collection should understand, however, that, regardless of the prior use indicated, it is unwise to take the questions and problems exactly as they are articulated in this book and in Part I and pass them, unaltered and unrefined, to their students. The items that are presented here are meant to be illustrative—to serve as nuclei, to provide hints and ideas for teachers to view as starting points for transformation to their own framework (or for entirely new questions of their own) rather than as fixed end points for immediate use. In other words, the suggested items should be reworked and reworded to fit smoothly and consistently with the presentations a teacher has used and, even more importantly, to fit the *vocabulary a teacher has been employing*. If these conditions are not met, questions, however well conceived in principle, will be meaningless and unintelligible to the students. The necessary transmutation must be effected by each individual teacher in the context in which he or she operates.

Because of this need for reworking and transmutation, and also to keep down the size and cost of this volume, I have not included my own answers and solutions. These are, in fact, irrelevant. Teachers should not make use

of questions about which they have doubts of any kind; they should rework any nucleus of an idea that appeals to them into a form that fits what they have been doing with their students. My own phraseology, for example, which may be clear to my own students because of the locutions I have used from the very beginning, may well be ambiguous in some other context and may require substantial revision for use by another teacher. Such revision, of necessity, remains up to the individual teacher.

It is my firm belief that testing along many of the lines illustrated in this collection (in addition to utilizing the more conventional end-of-chapter type problems) is essential for impelling students toward firmer concept formation and genuine understanding of the physics. In the absence of such testing, many students tend to memorize, without understanding, approaches to type-problem solving that yield sufficient partial credit for an adequate grade. They learn to avoid the labor of thought that yields the levels of understanding to which we (the teachers) persistently render lip service, and they emerge from our courses with the misapprehensions and lack of understanding so painfully revealed in the accumulating researches. Some of our students have never solved a problem completely correctly in their entire experience in physics, and we are vulnerable to the charge of encouraging the formation of what I have come to call “partial credit minds.”

In addition to questions and problems, I have included two other items that might be of use to some teachers. One is a set of statements of learning objectives that were formulated in our group at a time when we were operating summer institutes for high school physics teachers. (The participants were working in pairs in self-paced mode, and the statements of learning objectives were useful, even necessary, guides to help them achieve the depth of understanding we considered essential for teaching.) The other item consists of several examples of term paper assignments for an introductory physics course. We had found that term paper assignments were largely ineffective if beginning students departed into large subjects without some reasonable degree of constraint, guidance, and concentration. The examples given are intended to show how effective assignments, which provide guidance and choice without excessive constraint, can be constructed. They are intended as helpful models and suggestions rather than as items to be taken and implemented as they stand.

I am indebted to many sources for ideas entering into this collection. Some originated so long ago that I can no longer separate my own contribution from those of colleagues who worked with me over the years. I can only express gratitude to all who helped criticize and refine the questions that evolved in our courses. In more recent time, I am especially indebted to my colleague, the late Philip Peters, for a variety of fertile ideas and for his careful review of many items in this collection.

Contents of Part II

CHAPTER 1	Scaling and Ratio Reasoning	1
CHAPTER 2	Kinematics	11
CHAPTER 3	Force and Dynamics	26
CHAPTER 4	Momentum and Energy	50
CHAPTER 5	Electricity	71
CHAPTER 6	Direct Current Circuits	82
CHAPTER 7	Electromagnetism	95
CHAPTER 8	Particle Trajectories in E- and B-Fields	105
CHAPTER 9	Wave Phenomena	109
CHAPTER 10	Images with Mirrors and Lenses	122
CHAPTER 11	Geometrical and Physical Optics	130
CHAPTER 12	Fluids and Thermal Phenomena	139
CHAPTER 13	Kinetic Theory	155
CHAPTER 14	Modern Physics	159
CHAPTER 15	Mixed Areas of Subject Matter	170
CHAPTER 16	Naked Eye Astronomy	184
CHAPTER 17	Learning Objectives	193
CHAPTER 18	Term Paper Assignments	206

Chapter 1

Scaling and Ratio Reasoning

Note to the instructor: Most of the following questions are designed to lead students into thinking about scaling, and about functional relationships, in terms of ratios rather than through substitution in formulas. Probably because of lack of practice and unfamiliarity with this mode of thought, many students exhibit strong resistance to ratio reasoning and strive to convert it into formula substitution, avoiding the reasoning. Since such ratio reasoning, however, underlies virtually all estimating, scaling, and thinking in terms of orders of magnitude, its cultivation is highly desirable and important in student cognitive development. In this chapter, the range of context has been highly restricted, and the questions are essentially illustrative. Teachers would be well advised to provide many opportunities for such reasoning in every possible context of subject matter, using the basic pattern being illustrated if it fits the teacher's own mode of operation. To register their importance, however, it is essential to test on scaling and ratio reasoning. It is the usual *absence* of such testing that has left many students where they are and has helped crystallize their frequently massive, even fierce, resistance. The patterns illustrated in this chapter can be adapted to many areas of subject matter other than the ones here selected.

1.1 In a uniformly accelerated rectilinear motion, starting from rest, a particle undergoes a displacement of 3.36 m in a time interval $(\Delta t)_1$. What would have been the time interval $(\Delta t)_2$ for a displacement over the first 1.28 m in this same history? [Do not substitute in formulas. Select the relevant functional relationship between displacement and time interval and then express $(\Delta t)_2$ directly in terms of $(\Delta t)_1$ and an appropriate numerical ratio based on the functional relation.]

1.2 From the kinematic relations for uniformly accelerated rectilinear motion with negligible frictional resistance, find the expression for how high an object rises when it is thrown vertically upward with an initial velocity v_o . Explain your steps of reasoning in making the derivation.

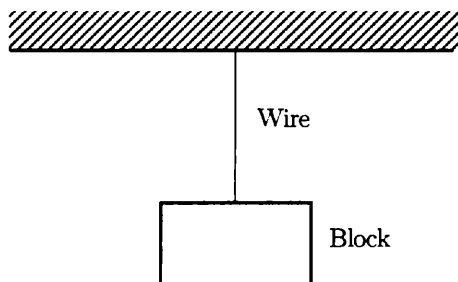
(a) The acceleration due to gravity at the surface of the moon is about $1/6$ that at the surface of the earth. For a given initial vertical velocity v_o , how much higher will the object rise if thrown upward on the moon than if thrown upward on the earth? Explain your reasoning. (Use *ratio reasoning* based on the functional relation you have obtained; do *not* substitute in the formula.)

(b) Suppose we increase the initial velocity by a factor of 2.6. By what factor will

the height of rise increase at each location? Explain your reasoning. (Use ratio reasoning based on the functional relation; do not substitute in the formula.)

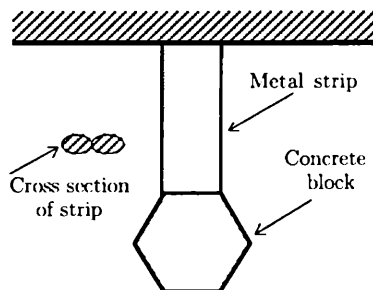
1.3 It is an established fact that the breaking strength of a stretched wire is directly proportional to the cross-sectional area of the wire. ["Breaking strength" is defined as the total stretching force (tension) at which the wire breaks.] Suppose a brass wire 0.56 mm in diameter is just approaching the breaking point under a tension of 85 N.

- (a) Suppose we wish to suspend a block with a mass of 24 kg on a brass wire as shown in the diagram. Explaining your reasoning, calculate the diameter of the wire that would be necessary to hold the block without breaking. (Pay careful attention to the units you use and to the distinction between weight and mass.)



- (b) Is the diameter you calculated in part (a) a minimum value, a maximum value, some intermediate value? Explain your reasoning.
- (c) Suppose the length of the suspending wire is now increased by a factor of 1.6 without any change in the cross-sectional area. Will the maximum weight the wire will support increase, decrease, or remain unchanged? Explain your reasoning.

1.4 A concrete block of the shape shown hangs from a thin metal strip having the cross-sectional shape indicated in the inset on the left of the diagram. Suppose the block is increased in size through scaling up each of its three linear dimensions by a factor of 2.10 without change in the material of which it is composed.



- (a) By what *factor* will the weight of the block be increased because of the change in dimensions? (Calculate the numerical value and explain your reasoning briefly.)
- (b) By what scale *factor* must the linear dimensions of the cross section of the suspending strip be changed to compensate for the change in weight of the block, making sure that the strip does not break? Explain your reasoning briefly.

(Remember that the breaking strength of the strip is proportional to its cross-sectional *area*, not its volume.)

1.5 Suppose that in a certain atmospheric situation, a falling raindrop, having an initial mass of 0.010 g, collects smaller droplets by collision at such a rate that, after 10 s, it has a mass of 0.020 g and a diameter of 3.4 mm.

- (a) If the droplet continues to gain mass at the same average rate as it continues to fall, what will be its diameter at the end of the next 10 s? (Use scaling and ratio reasoning, not substitution in formulas, and explain what you are doing.)
- (b) Is 3.4 mm a reasonable or unreasonable *order of magnitude* for the diameter of a droplet with a mass of 0.020 g? (Make a crude, quick estimate—not an elaborate calculation—and explain your reasoning.)

1.6 A statue is to be scaled down, without distorting its shape, by changing its total volume from 1.25 m^3 to 0.37 m^3 . Explain your reasoning in each of the following calculations.

- (a) If the height of the original statue is 250 cm, calculate the height of the smaller model.
- (b) If the circular base of the original statue has a circumference of 45 cm, calculate the circumference of the scaled-down base in the smaller model.
- (c) How will the total surface area of the model *compare* (this means an appropriate *ratio*) with the total surface area of the original? How will the surface areas of the circular bases compare?
- (d) If both the model and the original are made of the same material, how will the mass of the model compare with the mass of the original?
- (e) If the model and the original are not made of the same material, what would you have to know about the materials to be able to compare the masses, and how would you use this information?
- (f) If the original statue and the model turned out to have the same mass, what would you conclude about the materials making up the two objects? (Give a numerical answer comparing relevant properties of the materials.)

1.7 As a first approximation, let us think of an older person O as being a larger scale model of a younger person Y, with lengths in all parts of the body increased by the same scale factor. Suppose the weight of O is 4.50 times the weight of Y.

- (a) If the height of O is 6.0 ft, what must be the height of Y?
- (b) How will the cross-sectional area at some level in the legs of the older person compare with the cross-sectional area at the corresponding level in the legs of the younger?
- (c) Will the compressional stress on the leg bones be the same in the two individuals? (“Stress” is the name for the force per unit area that the bone must support.) Why or why not? If not, for which individual will the stress be

greater? By what numerical factor? It is because of the effect that emerges in the analysis you have just conducted that larger animals are, in reality in the world around us, not simply scaled-up models of smaller animals. Explain the reasoning behind this statement.

1.8 In explosions, the radius of a particular level of damage (say, the collapse of wooden buildings) varies as the cube root of the energy released in the explosion. (Energy release is frequently measured in terms of the equivalent number of pounds of TNT.) Suppose that in case A the energy release is equivalent to 200 lb of TNT while in case B the release is equivalent to 1200 lb.

- (a) How will the radii for a given level of damage compare in cases A and B? Explain your reasoning.
- (b) How will the area S_A in case A compare with the area S_B in case B for the same level of damage? Explain your reasoning.

1.9 A curve of radius R in a highway is banked at the optimum angle for cars traveling at a speed of 55 mi/h. By what factor must the radius of the curve be increased or decreased if the banking angle is to remain the same and still be the optimum value for trucks having a mass of 10,000 kg traveling at a speed of 40 mi/h? (This is a problem in ratio reasoning. Go to the equation that was derived for optimum angle of banking, establish the functional relation relevant to this problem, and calculate the scaling factor called for. Explain your reasoning, being careful to indicate what role is played by the mass and what role is played by the radius.)

1.10 The planet Saturn has a large moon called Titan. Titan has a mass 1.85 times the mass of our moon. Saturn itself has a mass 95 times that of the earth. Our moon has a mass 0.0123 times that of the earth. The distance between the centers of earth and moon is 240,000 miles, and the distance between centers of Saturn and Titan is 760,000 miles. Referring to the law of gravitation and explaining your reasoning in terms of scaling ratios *only* (without substitution in the formula), calculate the ratio of the centripetal force F_{TS} exerted on Titan by Saturn with the centripetal force F_{ME} exerted on the moon by the earth. (Give your argument in terms of whether F_{TS} will be larger or smaller than F_{ME} , and in what ratio, as prescribed by the gravitation law.)

1.11 The center of the moon executes a nearly circular orbit, with a radius of about 240,000 miles, around the center of the earth. Suppose the circumference of this orbit were increased by a length of 63 ft. (One mile contains 5280 ft.) By what amount would the earth-moon distance change? (Hint: To make the simplest and most efficient calculation, avoiding foolish complications with irrelevant numbers, sketch a graph of circumference versus radius for circles and locate the relevant circumference change, the corresponding radius change, and the relation between the two changes on your graph before plunging into calculations. If you find yourself working with huge numbers and converting between miles and feet, you are on a wrong track.)

Note to the instructor: In question 1.12, “density” is deliberately *not mentioned*. Use of this word directs many students to the formula for density and diverts them from the intrinsic arithmetical reasoning.

1.12 You have a block of wood with a total mass of 540 kg. This type of wood has 0.85 g in each cubic centimeter. Suppose you were to add 38 g of wood to the block. By how much would you increase the volume of the entire block? Do *not* substitute into a formula; explain the relevant arithmetical reasoning in your own words. (Hint: Be sure to think about *change* in volume rather than entire volume; look at the relation between the lines representing volume changes and corresponding mass changes on the graph of total mass versus total volume, and avoid making useless and irrelevant calculations.)

1.13 It is an empirical fact that the power output required of the engines of a boat or ship varies approximately as the cube of the speed; i.e., if you wish to double the speed of the vessel, you must increase the power output by a factor of 8.

- (a) Consider a boat with a mass of 2000 kg moving with initial speed v_i . The captain increases the power output of the engines by a factor of 2.6. By what factor does he increase the speed of the boat? Explain your reasoning.
- (b) By what factor does he increase the kinetic energy of the boat? Explain your reasoning.

1.14 Suppose a photographer has established the correct exposure time for a particular situation as determined by the brightness of the scene, the film being used, the lens, and the diameter of the lens opening. He now wishes to shift from his 50 mm focal length lens to a 120 mm lens and to reduce the exposure time by a factor of 3.6 while the light conditions and the film remain unchanged. Must he increase or decrease the diameter of the lens opening? By what numerical factor? Explain your reasoning.

1.15 It is established that oxygen atoms have 1.33 times the mass of carbon atoms. Suppose we have 100 g of oxygen and we want to weigh out an amount of carbon that has the same number of atoms as our sample of oxygen.

- (a) How many grams of carbon should we weigh out?
- (b) How would the number of atoms in 100 g of carbon compare with the number in 100 g of oxygen?

Explain your reasoning in both parts (a) and (b) in your own words. (Note that at this stage, we have no idea how many atoms of oxygen are actually present in 100 g of the gas, and we have no need for this information.)

1.16 Consider the case of interaction between two point charges A and B exerting a force of 0.0126 N on each other. Suppose that the magnitude of charge A is increased by a factor of 3.25, the magnitude of charge B is decreased to 0.873 its initial value, and the spacing between the charges is increased by a factor of 1.18.

- (a) Referring to Coulomb’s law to justify the ratio reasoning that is appropriate, calculate the final force acting on charge A by multiplying the initial force (0.0126 N) by numerical factors that increase or decrease the force in accordance

with the changes that are described. Do not substitute into the formula, but indicate the physical justification for each ratio applied. Put your calculation as a sequence along the following straight line:

$$0.0126 \times$$

1.17 Two point charges q_A and q_B are separated by a distance of 1.80 cm and interact with a force of 0.0035 N. Charge q_A is increased by a factor of 8.2 and the separation is increased to 4.30 cm. By what factor must charge q_B be changed if the force of interaction is to end up at a value of 0.0045 N? (Solve the problem directly by ratio reasoning without formal substitution into Coulomb's law.) Explain your reasoning.

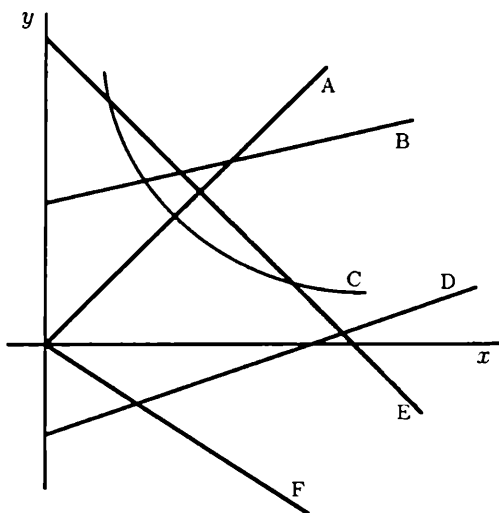
1.18 Two point charges A and B attract each other with a particular value of force. Suppose that the magnitude of charge A is increased by a factor of 5.3 while that of B is decreased by a factor of 2.4. By what factor must the spacing between the charges be changed to keep the force of interaction unchanged from its initial value? Be sure to explain your ratio reasoning and to indicate whether the spacing is to be increased or decreased.

1.19 In the electrolysis of water, the passage of an electrical charge of 96,500 C liberates 1.008 g of hydrogen. From its chemical properties, we know that the hydrogen ion is associated with one corpuscle of electrical charge, i.e., is represented by the symbol H^+ . It is known that the nitrogen atom (N) has a relative mass of 14.0 on the scale in which hydrogen has the relative mass 1.008. It is also known that ordinary nitrogen gas is diatomic; i.e., it has the molecular formula N_2 .

- Calculate the charge-to-mass ratio for H^+ in coulombs per kilogram.
- Suppose you are looking for doubly ionized nitrogen molecules N_2^{2+} in a beam of positive ions. Reasoning in terms of available ratios, calculate the expected charge-to-mass ratio of N_2^{2+} in coulombs per kilogram, starting with the value for H^+ and using ratio reasoning. Explain your reasoning.

1.20 Consider the following graphs:

- Which, if any, of the graphs represent direct proportions?
- Which, if any, of the graphs represent linear relations that are *not* direct proportions?
- Which, if any, of the graphs represent nonlinear relations?
- Which, if any, of the graphs might represent inverse proportions?



1.21 Consider the following names for different functional relationships: (1) “direct proportionality,” (2) “linear relation but not a direct proportionality,” (3) “nonlinear relation,” (4) “inverse proportionality,” (5) “inverse square relation.”

- (a) Sketch a graph illustrating each relationship on a set of coordinate (x and y) axes.
- (b) Write an algebraic equation in terms of y and x corresponding to each illustration.
- (c) Illustrate each type of relation by describing a physical situation in which you have encountered it.

1.22 Consider the following argument between two individuals who are comparing two bodies, A and B, of possibly different composition by comparing their densities. Person W argues that since the mass of B is twice that of A, the density of B must be twice that of A because density is directly proportional to mass. Person Z, on the other hand, argues that since the volume of B is twice that of A, the density of B must be half that of A because density is inversely proportional to volume.

- (a) Assuming the information about masses and volumes is correct, what is the actual relation between the two densities? Explain your reasoning.
- (b) Describe in your own words just where the fallacy in the arguments presented by W and Z lies.

1.23 Centripetal acceleration of an object in uniform circular motion can be expressed in two alternative forms:

$$a_c = v^2/r \quad \text{or} \quad a_c = r\omega^2$$

where r denotes the radius of the circle, v the magnitude of the tangential velocity of the object, and ω its angular velocity. Since $v = r\omega$, it is clear that the two expressions are equivalent.

Using the first expression, person A argues that since centripetal acceleration is inversely proportional to r , centripetal acceleration of the object must be larger in smaller circles than in larger ones. Person B disagrees and, appealing to the second expression, argues that, since centripetal acceleration is directly proportional to r , centripetal acceleration of the object must be smaller in smaller circles.

- (a) Starting with the definitions of v and ω and with the help of a simple diagram, explain where the relation $v = r\omega$ comes from.
- (b) Identify the source of the disagreement between A and B in the foregoing arguments, and alter each statement to make it correct and eliminate the inconsistency between the two.

1.24 Consider the following two equations:

$$(1) y = 3x - 5 \quad \text{and} \quad (2) \bar{v} = \Delta s / \Delta t$$

where the latter refers to the concept of average velocity in kinematics.

- (a) Describe in your own words the difference between statements (1) and (2) as far as their origin, nature, and logical content are concerned. [Hint: How did (2) originate? Why would it be more appropriate in (2) to use the symbol \equiv (meaning “defined as” or “identical to”) rather than the ordinary $=$ sign?]

Consider two additional equations:

$$(3) (x + y)^2 = x^2 + 2xy + y^2 \quad \text{and} \quad (4) \cos 60^\circ = 0.500$$

- (b) What differences do you discern among the meanings of the equals signs in equations (1), (3), and (4)? Is (3) a functional relation in the same sense as (1)? Why or why not? Is (4) a functional relation? Describe the differences in detail in your own words, with appropriate illustrations or examples.

Consider the two equations:

$$(5) 2y = 5x \quad \text{and} \quad (6) 2 \text{ in} = 5 \text{ cm}$$

where (5) is an ordinary algebraic relation and (6) is the relation between the different measures of length (i.e., inches and centimeters).

- (c) Indicate the ways in which the statements (5) and (6) are profoundly different in meaning even though the same symbol ($=$) is (somewhat misleadingly) used in both. [Hint: Is 2 times the number of inches equal to 5 times the number of centimeters? (That is, are “in” and “cm” in (6) similar in meaning to the symbols y and x in (5)?)]
- (d) What about equations such as $\vec{F}_{\text{net}} = m\vec{a}$ or $\vec{F} = Gm_1m_2/r^2$? Does the equals sign have the same meaning it does in the preceding examples? (Do not forget the extended “stories” and definitions that go with generating the meaning of these two relations.)
- (e) Does the equals sign always have exactly the same meaning wherever it happens to arise in your textbooks? Explain to a fellow student who is having trouble seeing the point of this entire question what it tells you about the caution you should exercise in connection with casual textbook usage of the same symbol ($=$) in very different contexts and physical situations.

1.25 You have probably seen a film in which a flower, starting as a bud, opens up before your eyes. This speeding-up effect is accomplished by “time lapse photography,” in which the frames are snapshots, taken with a carefully chosen time interval between them and not as a continuous “motion picture.” In this problem you are asked to make your own estimates and judgments as to the time interval between frames that will yield the desired speeding-up effect. (Estimating is not a matter of wild guesswork without underpinnings. It is a matter of careful reasoning with meaningful, albeit not highly precise, values.)

- (a) Suppose you want to present on the screen the complete opening of a flower (daffodil? rhododendron? camellia? anything you like) from the initial bud to the fully opened flower. How long do you want the entire sequence to last when it is finally projected?

- (b) In films, the illusion of motion is produced by flashing successive pictures on to the screen at the rate of about 20 frames per second. How many pictures will you need to present your entire speeded-up version of the opening of the flower? About how long does it take for the flower to open? (You must make your own reasonable estimates of relevant numerical values.)
- (c) Time lapse photography is carried out by focussing on a given object a movie camera equipped with an automatic triggering device that permits successive snapshots to be taken on successive frames of film at intervals of seconds, minutes, hours—whatever is desired—instead of running the film continuously at high speed. To obtain your movie of the flower, what time lapse (i.e., what time interval between successive frames) will you want to set? Carry out the calculation and explain your reasoning.
- (d) Now explain to a novice, invoking numerical data and not just words, why you would not want to photograph the sequence at normal camera speed and then run it very much faster to show the flowering in “slow motion.”

1.26 The speed of light is known to be about 186,000 mi/s. Use denary (power of ten) notation in making the following calculations, and explain your reasoning throughout. Do not use formulas in which to substitute. Present your solutions in terms of simple arithmetic and purely arithmetical reasoning. Do not give numerical answers to any more than the justifiable number of significant figures.

- (a) Light from the sun requires about 8 minutes to travel from the sun to the earth. How far from earth is the sun (in miles)?
- (b) The average distance from the earth to the moon is about 240,000 miles. How long does it take a flash of laser light to travel from the earth to a target on the moon?
- (c) Read the following question regarding mass and volume for understanding, but do not answer it.

Metallic copper contains a mass of 8.9 g in each cubic centimeter. The volume of a copper bar is known to be 50 cm^3 . What is the mass of the bar?

Here is the question to be answered: Is the reasoning to be used in the question regarding the copper more like that used in part (a) or in part (b)? Explain your answer.

- (d) Make up a problem about the metallic copper in which the reasoning to be used would be analogous to whichever [either (a) or (b)] did not apply to (c).

1.27 It is an experimental fact that the total electrical resistance of any metallic wire is directly proportional to its length and inversely proportional to its cross-sectional area (in other words, inversely proportional to the square of the diameter). The proportionality constant is, of course, different for each different metal and is thus a property of any given metal.

- (a) In your own words, argue that these experimental facts indicate that in the metallic wire, electric charge must be moving throughout the entire cross section of the wire and is not, for example, confined to the surface. If the current were confined to the surface, what would you expect to have been the relation between resistance and wire diameter? Explain your reasoning.

(This is an example of how important it sometimes is to recognize what is *not* the case and contrast it with what *is* the case. The resistance might conceivably have been inversely proportional to the diameter but, in fact, it is not. It is an experimental fact, however, that if current in the wire is alternated at extremely high frequency, the motion of charge *is* confined to the surface. This so-called “skin effect” does not play a role in circuits with which we are concerned in introductory physics.)

- (b) Suppose a solid rod of metal has a radius a , a length L , and a resistance R . A hollow rod is now made of the same material with the same length and the same outer radius, but its inner radius is half the outer radius. How will the resistance of the hollow rod compare with that of the solid rod? Obtain the numerical value of the ratio and explain your reasoning.

1.28 A point on the earth’s equator has a tangential velocity v_{rot} (relative to the fixed stars) by virtue of rotation about its own axis. The earth also has a tangential velocity v_{rev} in its revolution about the sun. The radius of the earth’s solar orbit is about 23,400 times as large as the earth’s diameter.

- (a) Using your everyday knowledge of relevant time intervals (i.e., length of day and length of year), compare the two tangential velocities. Which one is larger? By what numerical factor? Use ratio reasoning without substitution in formulas and explain each step.
- (b) How do the tangential velocities of rotation at points at latitude 45°N or S and at the north pole or south pole compare with the tangential velocity of revolution around the sun? Draw an appropriate diagram to show the tangential velocities at different latitudes and explain your reasoning in obtaining the various numerical ratios.
- (c) [For students who might have studied the Michelson-Morley experiment: What relevance do these comparisons have to the Michelson-Morley experiment?]

1.29 A car starts from rest at position $s = 0$ and accelerates uniformly along a straight road in the positive s direction. At position s_1 it has a velocity v_1 . What will be its velocity at positions $2s_1$ and $3s_1$? Express your result in terms of v_1 multiplied by a numerical factor and explain your reasoning. (Do your reasoning in terms of the ratios called for by the applicable functional relation, not by substitution in a formula.)

1.30 A ball rolls down an inclined plane of length L starting from rest at the top. Where along the plane would it have half the velocity it reaches at the bottom? Use ratio reasoning based on the relevant functional relation; do not just substitute into a formula. Explain your reasoning.

Chapter 2

Kinematics

2.1 In cases of rectilinear motion, we give the name “average velocity” and the symbol \bar{v} to the quantity $\Delta s/\Delta t$, where s represents position numbers along the straight line in question and t represents clock readings at corresponding values of s . \bar{v} is a single number characterizing the motion during the time interval Δt regardless of how complex and variable the velocity history may have been. It is a vector quantity in the sense that it is accompanied by plus or minus signs depending on the algebraic sign of Δs . The symbol $|\bar{v}|$ would be described as the “magnitude of the average velocity for the given time interval,” meaning the size of this quantity regardless of its algebraic sign (direction).

- (a) Consider the quantity that would be described as the “average magnitude of the instantaneous velocity” and would be represented by the symbol $|\bar{v}|$. How would you go about calculating $|\bar{v}|$? Describe the process in detail.
- (b) Describe some motions in which the two quantities $|\bar{v}|$ and $|\bar{v}|$ would come out equal. Describe some motions for which they would *not* be equal. Include in your discussion the case in which the body, at the end to the interval Δt , returns to the same position it occupied at the beginning.

2.2 If we are dealing with a situation in which the instantaneous velocity v changes significantly over the given time interval Δt , what is wrong with saying $\Delta s = v\Delta t$? What is the only value of velocity that will make this equation correct under these circumstances? Explain your reasoning.

2.3 Consider making the calculation $\Delta t/\Delta s$ instead of $\Delta s/\Delta t$. Is this just a piece of foolishness, or does the number so obtained have a reasonable and intelligible physical meaning?

- (a) Interpret $\Delta t/\Delta s$ in words, noting that the interpretation, in words, of $\Delta s/\Delta t$ is “the number of meters change of position taking place in *one* second.”
- (b) Execute with movements of your hand (1) a motion for which $\Delta t/\Delta s$ would have a very high value and (2) a motion for which it would have a very low value, and contrast your actions with those you would execute if illustrating $\Delta s/\Delta t$.

- (c) Invent a good, descriptive name to give the quantity defined by $\Delta t/\Delta s$. (Geophysicists actually do have a descriptive name for this quantity; see if you can guess what it is in the light of the character of the motion when $\Delta t/\Delta s$ is very large.)

2.4 In this question, we shall do some qualitative reasoning concerning the effect of air resistance on the motion of a ball thrown vertically upward with initial velocity v_{0y} , rising to its maximum height, and then falling back to the level from which it was thrown. (The velocity of the thrown ball is great enough for air resistance to play a significant role but not so great that the ball approaches terminal velocity at any time during its flight.) We raise the following question: How does the time interval $(\Delta t)_{\text{up}}$ for rise to maximum height compare with the time interval $(\Delta t)_{\text{down}}$ for falling from maximum height back to the starting level? Record your initial intuitive response to this question: Are the two time intervals equal or unequal? If unequal, which is greater? Explain the basis for your intuitive response.

Now let us analyze the situation through a careful sequence of questions, making use of the following notation. Under conditions of negligible air resistance, the magnitude of the vertical acceleration is denoted by g . Let us denote the magnitude of the average acceleration during the upward motion in the presence of air resistance by $a_{\text{u}y}$ and the magnitude of the average acceleration during the downward motion by $a_{\text{d}y}$.

- How would you expect the height of rise in the presence of significant air resistance to compare with the height in the case of negligible air resistance? Explain your reasoning by referring to the relevant kinematic equation.
- How would you expect the magnitude of $a_{\text{u}y}$ to compare with g ? Would it be greater than, equal to, or smaller than g ? Explain your reasoning. Is your answer consistent with the answer you gave in part (a)? Explain.
- How would you expect the magnitude of $a_{\text{d}y}$ to compare with g ? Explain your reasoning.
- In the light of your answers in parts (b) and (c), how will the time interval $(\Delta t)_{\text{up}}$ for rising to maximum height compare with the time interval $(\Delta t)_{\text{down}}$ for falling from maximum height back to the starting level? Explain your reasoning. Is the result you have arrived at consistent with the one you anticipated intuitively at the start of the question? If not, reexamine what you have done, with the objective of achieving internal consistency. (Hint: It helps to keep in mind that the distances of rise and fall are the same.)

2.5 A person standing at the edge of the roof of a building throws ball A vertically upward with an initial velocity $+|v_{0y}|$ and throws a second ball B vertically downward with an initial velocity $-|v_{0y}|$ of the same magnitude. Both balls fall to the ground past the edge of the building.

- Under conditions of negligible air resistance, what is the velocity of ball A as it passes, on the way down, the level from which it was thrown? Explain your reasoning.
- Under conditions of negligible air resistance, how will the velocities of the two balls on striking the ground compare with each other? Explain your reasoning.

- (c) Now let us visualize the same experiment performed in the presence of significant air resistance. How will the two velocities on striking the ground compare with each other? Will they be equal in magnitude or unequal? If unequal, which will be greater and why?
- (d) Suppose that a third ball C is projected from the edge of the roof with the same vertical component of velocity $+|v_{0y}|$ but with a horizontal component of the same magnitude. In the case of negligible air resistance, how will the magnitude of the *total* velocity of C on striking the ground compare with that of A? (Draw a relevant diagram; find the numerical value of the ratio; explain your reasoning.)

2.6 You have an ordinary stopwatch like that used in timing athletic events. Describe how you might take advantage of the relation $\Delta s = (1/2)g(\Delta t)^2$ and the known value of g to determine the height of a window above the ground (or the height of a bridge above a stream) by dropping an object from the upper location. Examine the accuracy to be expected under various circumstances: What trouble would you run into if the height is relatively small? How small is “relatively small”? About how large would the height have to be for you to obtain reasonably reliable values? What troubles do you begin to run into as the height becomes quite large?

2.7 Consider a car moving along a highway. Answer the following questions, giving an explanation of your answer in each case. Sketch at least one possible set of a versus t , v versus t , and s versus t graphs corresponding to your description in each case.

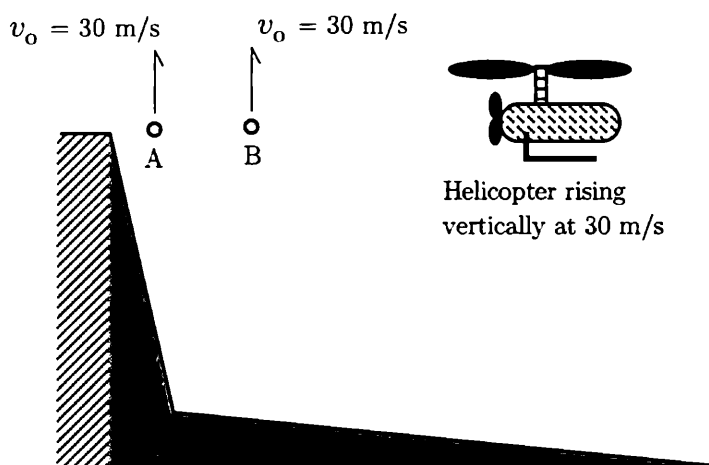
- (a) If the acceleration a of the car is zero, what are the possible values of velocity?
- (b) If the car is moving, is it necessarily accelerating?
- (c) If the car is not accelerating, is it necessarily standing still?
- (d) Under what circumstances is the acceleration in the opposite direction to the velocity?
- (e) How might you drive a car so that the acceleration would go through zero (from positive to negative) while the velocity remains positive?
- (f) How might you drive a car so that the velocity would go through zero (from negative to positive) while the acceleration remains positive?

2.8 Suppose you are driving a car and are accelerating, increasing your speed in the positive direction. You now relax slowly on the gas pedal, decreasing the *magnitude* of your acceleration (you do *not* use the brake).

- (a) Are you increasing or decreasing your speed as the *magnitude* of the acceleration decreases? Explain your reasoning.
- (b) Sketch a versus t and v versus t graphs for the situation under consideration and make sure that your graphs are consistent with the verbal description you gave in part (a).
- (c) Sketch corresponding graphs for the same sequence except for initial motion in the negative direction.

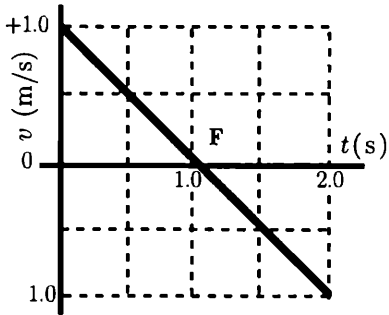
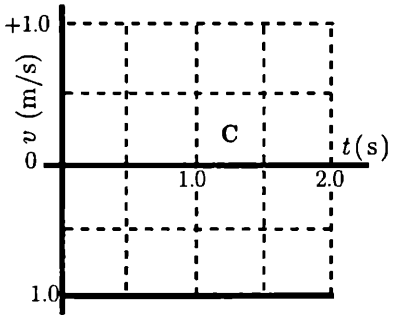
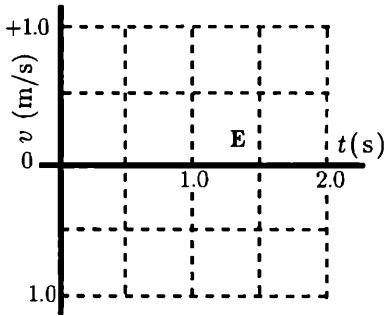
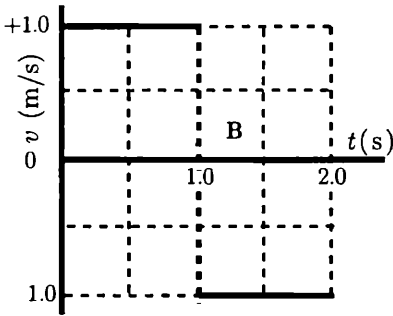
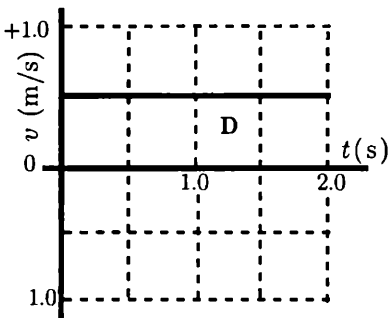
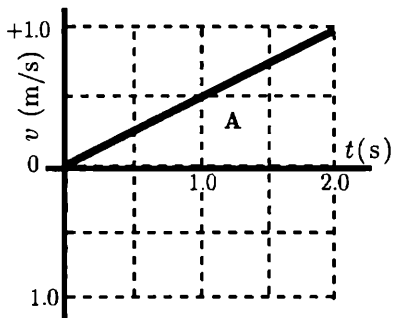
2.9 A numerical value of acceleration can be interpreted as telling us “how fast the velocity of a body is changing.” Starting with the definition of acceleration, explain why this is a legitimate statement. Now consider the following statement: “A numerical value of acceleration, given alone, tells us nothing about how fast the object in question is moving.” Is this statement correct or incorrect? Explain your answer carefully by referring to the definitions of both velocity and acceleration.

2.10 Consider the situation sketched in the figure: at clock reading $t = 0$, observer A at the edge of the cliff throws a ball vertically upward with an initial velocity of 30 m/s. Observer B is located in the helicopter, which started at the base of the cliff and is rising vertically at the constant velocity of 30 m/s. At the same clock reading $t = 0$ at which A throws his ball, B is passing A and releases a ball from his window. (B simply lets go the ball without any throwing action.) B continues upward in the helicopter with no change in the upward velocity of 30 m/s.



- Describe how observer A perceives the motion of the two balls relative to his position on the edge of the cliff after the instant $t = 0$. That is, what does each ball do relative to this observer? How does A describe the velocity as varying? How does A describe the acceleration? Cover the entire sequence between $t = 0$ and the instant the balls finally land at the base of the cliff.
- Describe how observer B perceives the motion of the two balls, not relative to the ground but relative to B's frame of reference in the rising helicopter. In particular, what is each ball doing relative to B at the instant A claims that the ball has reached the top of its flight? How does B describe the velocity as varying? How does B describe the acceleration?

2.11 The following velocity versus clock reading histories describe the rectilinear motion of six particles that started out from position s_0 at $t = 0$ s. Circle the correct answers for each of the following questions.



(a) Which particle (or particles) have returned to position s_0 at the clock reading $t = 2.0$ s?

A B C D E F NONE

(b) Which particle (or particles) spend at least some time moving in the negative direction?

A B C D E F NONE

(c) Which particle (or particles) move at uniform, nonzero acceleration?

A B C D E F NONE

- (d) Which particle (or particles) started in the negative s direction and then reversed the direction of motion, traveling back in the positive s direction?

A B C D E F NONE

- (e) Which particle is farthest from position s_0 at clock reading $t = 2.0$ s?

A B C D E F NONE

- (f) Which particles are the same distance from s_0 at $t = 2.0$ s?

A B C D E F NONE

- (g) Which particle (or particles) exhibited nonzero acceleration during the given period?

A B C D E F NONE

- (h) Which particle (or particles) kept moving in the same direction throughout the given period?

A B C D E F NONE

- (i) Which particle (or particles) exhibited negative acceleration over some interval?

A B C D E F NONE

- (j) Which particle (or particles) stood still for some time at some point in the history?

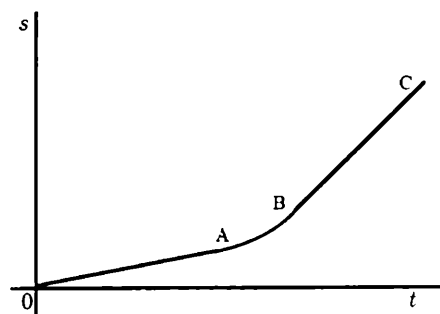
A B C D E F NONE

- (k) Which particle (or particles) move with nonuniform, increasing acceleration?

A B C D E F NONE

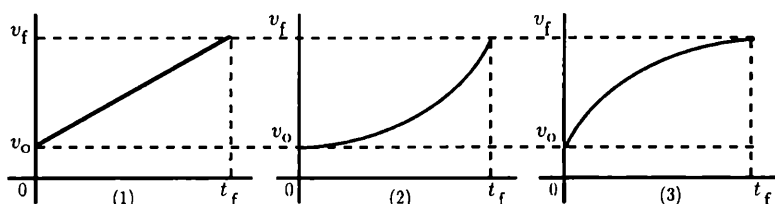
2.12 The diagram shows the position s versus clock reading t history of the motion of a car starting at $t = 0$.

Describe in your own words what you would see the speedometer needle doing during the various portions of the history: from 0 to A, from A to B, from B to C. Does any acceleration take place? If so, over what interval?



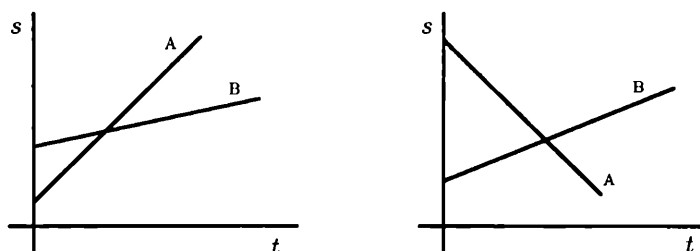
2.13 Consider the three different histories of velocity v versus clock reading t shown in the figure.

- (a) In case 1, argue that the average velocity \bar{v} over the time interval between $t = 0$ and $t = t_f$ must be equal to $(v_o + v_f)/2$. Explain your reasoning carefully, making use of the fact that the history is a straight line.



- (b) In cases 2 and 3, is \bar{v} also equal to $(v_o + v_f)/2$? Why or why not? If not, how does \bar{v} compare with $(v_o + v_f)/2$ in each instance? Is it greater or smaller? Explain your reasoning.

2.14 The graphs show the position-clock reading histories of the simultaneous motions of two cars A and B in parallel lanes along a straight highway in two different occurrences.

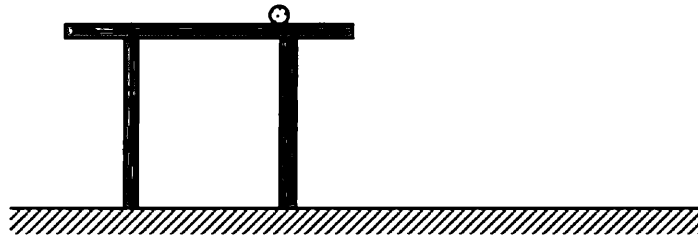


- (a) Are there any instants at which the two cars have the same velocity? If so, mark the instant (or instants) along the t -axis with the symbol t_{eq} and explain how you arrived at your conclusion. If not, explain how you arrived at that conclusion.
- (b) Are there any instants at which the two cars are passing each other, i.e., are at the same distance from the origin of position s ? If so, mark the instant or instants with the symbol t_p . Explain your reasoning. In each case, which car is going faster at the instant of passing? Explain your reasoning.
- (c) At the instant the cars are located at the same position s , have they traveled the same or different distances from their own initial positions at $t = 0$? If the distances are unequal, indicate which is larger. Explain your reasoning.
- (d) Suppose the speed of car B is increased somewhat in each case. What will happen to the clock reading at which the two cars will be located at the same position? What will happen to the distance each car will have travelled from its $t = 0$ position? Explain your reasoning by altering the diagrams to show what happens in each case.

2.15 Suppose that you are driving in car A along a straight road at 10 mi/h. A friend in car B is driving at a constant velocity of 60 mi/h (88 ft/s). At clock reading $t = 0$ and position $x = 0$, car B passes you and continues without change in velocity. You, however, in car A, step on the accelerator and maintain a constant acceleration of 5.0 ft/(s)(s).

- (a) Sketch a *qualitative* (do not try to plot numerically) graph of the x versus t history of the motion of the two cars. (b) On the basis of the diagram you have sketched, infer whether or not you will ever overtake car B. Explain your reasoning.
- (c) If, in part (b), you concluded that you will overtake B, calculate the clock reading at which the overtaking will occur and calculate the velocity (in both mi/h and ft/s) you will have attained at the instant of overtaking. Be sure to examine and interpret your results to determine whether or not the plan for overtaking is realistic.

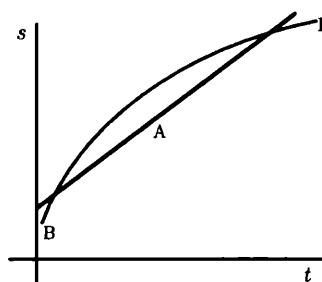
2.16 A ball is fired off the edge of a table with a horizontal velocity v_x and lands on the floor.



- (a) On the diagram, sketch a possible trajectory (the path followed by the ball) from the edge of the table to the floor.
- (b) Now, on the same diagram, sketch two other trajectories, one for a smaller value of v_x and one for a somewhat larger value of v_x . Sketch a fourth trajectory for an extremely large value of v_x . Label each of the trajectories with the comparative sizes of v_x .
- (c) We say that the shape of such trajectories is parabolic in the ideal case in which air friction is negligible. What is a “parabola”? That is, how is this kind of curve defined? How do we know that the shape of the trajectory is parabolic?
- (d) For the drop of the ball from its original level at the height of the table to be virtually unobservable to the naked eye at the wall of the room some reasonable distance from the table, what would have to be the numerical magnitude of v_x ? (You will have to choose your own reasonable values for the drop and for the distance to the wall.) What effect would you expect the ball to have on the wall under these circumstances? (Justify your answer.)
- (e) Return to part (a) and sketch another trajectory: That of another ball having a very much larger mass than that of the first ball but exactly the same initial velocity v_x . Explain your reasoning.

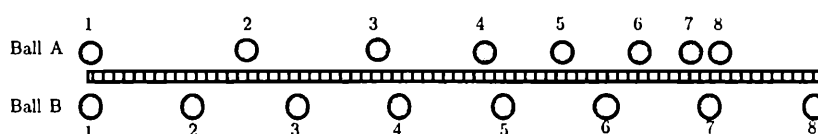
2.17 The figure shows position versus clock reading histories of rectilinear motions of two balls A and B rolling on parallel tracks.

- Mark with the symbol t_p along the t -axis on the diagram any instant or instants at which one ball is passing the other.
- Which ball, A or B, is moving faster at any of the clock readings t_p ?
- Mark with the symbol t_{eq} along the t -axis any instant or instants at which the two balls have the same velocity.



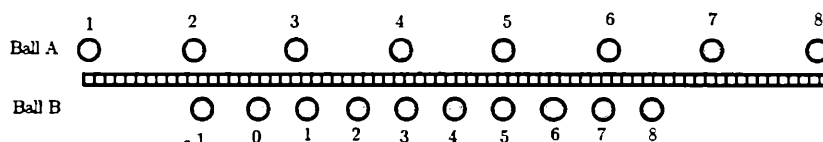
- Circle the correct statement from among the following: Over the period of time shown in the diagram, ball B is
 - speeding up all the time.
 - slowing down all the time.
 - speeding up part of the time and slowing down part of the time.
 - neither speeding up nor slowing down.
- Over the time interval between the passing points, does ball B travel a greater distance, a smaller distance, or the same distance as ball A?

2.18 The figure represents a flash (or stroboscopic) photograph looking down on two balls rolling parallel to a position scale on a level table top. The numbers show the clock readings corresponding to each ball position.



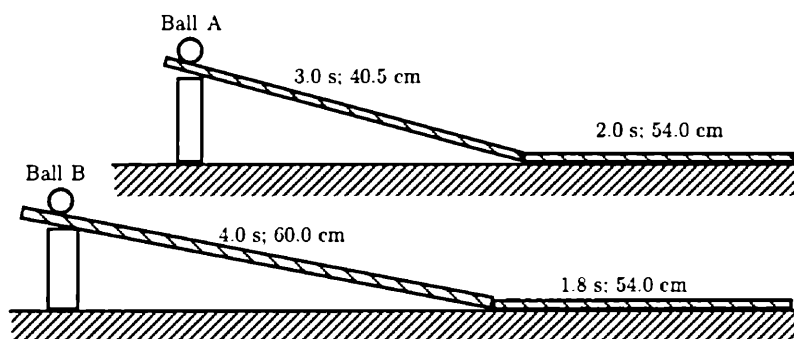
- At approximately what clock reading (or readings) do the two balls have very nearly the same speed?
- At approximately what clock reading (or readings) does one ball pass the other? In each instance you cite, indicate which ball, A or B, is doing the overtaking.
- Sketch position versus clock reading graphs for each of the two balls on the same set of axes, showing clearly how the two motions are related. Be sure to check whether your answers in parts (a) and (b) are consistent with your graphs. Explain your reasoning.

2.19 In the figure, we are looking down on a level table top. Assume that a flash (or stroboscopic) photograph has been taken of two balls A and B rolling parallel to a position scale on the table. Corresponding clock readings are shown at each image.



- At what clock readings, if any, do the two balls have very nearly the same speed?
- At what instant or instants, if any, is ball B overtaking and passing ball A?
- Sketch position versus clock reading graphs for each of the two balls on the same set of axes, showing clearly how the two motions are related. Check whether your answers to parts (a) and (b) are consistent with your graphs, and explain your reasoning.

2.20 Two balls A and B are released from rest and roll down sloping sections of track as shown. The slopes of the two tracks are not necessarily the same as those in the diagram. At the foot of each slope, the balls roll on to level sections of track along which they continue at uniform velocity. Certain measured times and distances are shown.



- According to the information given, which ball has the greater acceleration on its sloping section of track? Base your analysis directly on the fundamental definition of acceleration. No credit will be given if you use derived kinematic relations. Show all your calculations and explain your reasoning.

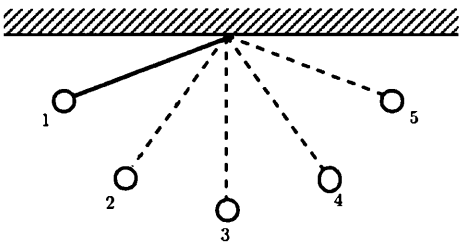
2.21 You are called upon to explain to a fellow student how it comes about that the change in position Δs for an object in rectilinear motion during a time interval Δt might carry either a positive or a negative sign.

- Present your explanation, making use of appropriate sketches or diagrams and connecting your description with some easily visualizable situation, such as moving in a car or watching a cart on the table.
- Making use of the explanation you have given in part (a) and the definition of average velocity, explain to a fellow student how it comes about that velocities might be either positive or negative. Explain how the algebraic signs must be

- interpreted, connecting your explanations with the diagrams you utilized in part (a).
- (c) Making use of the definition of acceleration and the explanations you gave in parts (a) and (b), explain to a fellow student how it comes about that acceleration values might come out either positive or negative. Explain how the algebraic signs must be interpreted, connecting your explanations with the experiences and diagrams you utilized earlier.

2.22 A pendulum bob, released from rest at position 1, can swing to position 5.

Suppose that as the bob swings, the string is suddenly cut at position 2, or position 3, or position 4, or position 5. In each instance, sketch the trajectory the bob would follow if the string were cut at the instant the moving bob reached that position, and explain your reasoning.



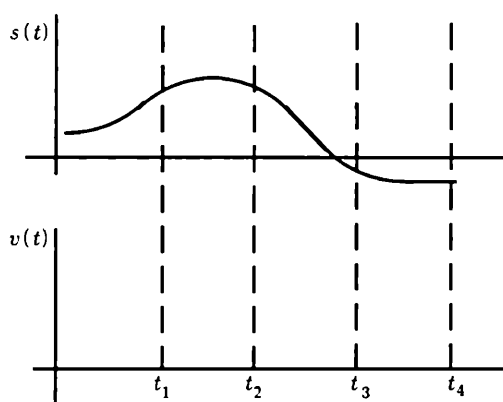
2.23 The table shows histories of instantaneous velocity versus clock readings for two different cases in an accelerating car.

Clock and Speedometer Readings (Instantaneous Velocities) in an Accelerating Car				
	Case I		Case II	
Clock reading t (s)	Speedometer reading v (mi/hr)	Average acceleration \bar{a} [mi/(hr)(s)]	Speedometer reading v (mi/hr)	Average acceleration \bar{a} [mi/(hr)(s)]
0.0	+6.0		+20.0	
1.0	+4.0		+10.0	
2.0	+2.0		+ 5.0	
3.0	0.0		+ 1.0	
4.0	-2.0		0.0	
5.0	-4.0		0.0	

- (a) In the blank columns, enter values of average acceleration \bar{a} corresponding to each time interval for case I and case II.
- (b) Sketch (do not try to plot) a versus t , v versus t , and position s versus t diagrams for each of the two cases.
- (c) Describe how, if your were driving, you might actually make a car execute (approximately) each one of the two motions.
- 2.24 Consider the following data on the rectilinear motion of a car that starts from rest at clock reading $t = 0$ and position $x = 0$. At clock reading $t = 5.0\text{s}$, it is observed to be at position $x = +40.0\text{m}$ and to have an instantaneous velocity of $+11.0\text{ m/s}$.

- Examine the interconnections among the given data carefully. Was the acceleration of the car uniform? Explain your reasoning.
- Are the kinematic equations such as $v = v_o + at$ and $x = (1/2)at^2$ applicable throughout the history of the motion? Why or why not?
- Sketch the shape of the v versus t graph that is implied by the data: Is the graph straight or curved? If curved, is it concave up or concave down? Explain your reasoning.

2.25 The figure shows a schematic position s versus clock reading t history of the rectilinear motion of a body.



- On the $v(t)$ versus t coordinates, sketch the corresponding velocity versus clock reading history.
- The number $\int_{t_1}^{t_2} v(t) dt$ has a geometrical interpretation on the $v(t)$ versus t diagram. Indicate on this diagram what this interpretation is.
- The number $\int_{t_1}^{t_2} v(t) dt$ also has a geometrical interpretation on the $s(t)$ versus t diagram. Indicate on this diagram what this interpretation is.

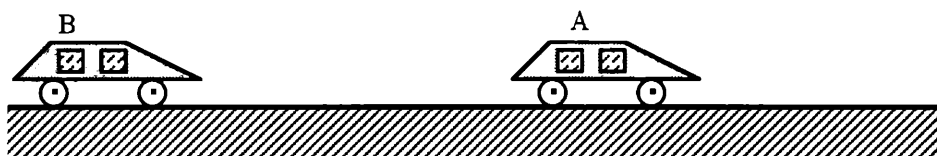
2.26 A cart rolls down an inclined plane starting from rest at the top. It is halfway down after a time interval Δt . Will the time interval to get the rest of the way to the bottom be equal to, greater than, or less than Δt ? Explain your reasoning.

2.27 A car starts from rest at position $s = 0$ and moves in the positive s direction. At position s_1 it is observed to have a velocity v_1 . Some time later, at position $2s_1$, it is observed to have a velocity $2v_1$. At position $3s_1$, it has a velocity $3v_1$, etc. Is the car accelerating? If it is accelerating, is the acceleration uniform, increasing, or decreasing? Explain your reasoning.

2.28 Over a certain time interval Δt , a car moves in such a way that the magnitude of its instantaneous velocity is never *smaller* than the magnitude of the average velocity.

- (a) Sketch a possible s versus t graph and comment on whether the car could have been accelerating during the given interval. Examine, in the same way, the case in which the magnitude of the instantaneous velocity is never *larger* than the magnitude of the average velocity.
- (b) Sketch s versus t graphs for two cases: One in which the car is speeding up uniformly and the other in which it is slowing down uniformly over the same time interval Δt , making the average velocity the same in both graphs. Making use of the graphs, comment on how instantaneous velocities must compare with the average velocity at various instants during each of the two histories: Must there always be instantaneous velocities both greater and smaller than the average? Why or why not? Must there always be an instantaneous velocity that is equal to the average velocity? Why or why not?
- (c) For those who have studied calculus: Can you connect the graphical and intuitive observations you make in parts (a) and (b) with any general theorems you developed in the calculus?

2.29 Consider a case in which two cars, A and B, on a straight road are located one behind the other as shown.



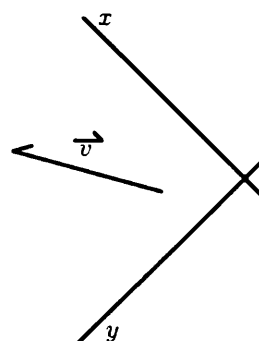
- (a) Suppose car A starts from rest with uniform acceleration a at instant $t = 0$, moving in the positive direction (to the right) along the road. Car B starts somewhat later, at instant $t = t_1$, with exactly the same acceleration. What will happen to the spacing between the two cars as time goes by (while they are still accelerating)? Will they remain the same distance apart? Will the spacing keep decreasing? Will the spacing keep increasing? Explain your reasoning and support it both with relevant diagrams and with an algebraic analysis.
- (b) Suppose we invert the situation as follows: Car A has an instantaneous velocity v_0 to the right at $t = 0$. Car B, somewhat behind car A, has the same instantaneous velocity at a somewhat later instant $t = t_1$. Both cars *slow down* uniformly with the same negative acceleration a . What will happen to the spacing between the two cars as time goes by (while they are still moving toward the right)? Will they remain the same distance apart? Will the spacing keep decreasing? Will the spacing keep increasing? Explain your reasoning and support it both with relevant diagrams and with an algebraic analysis.

2.30 A cart, released from rest at the top of an inclined plane, rolls down the plane, striking a spring at the bottom. It compresses the spring to the point at which its instantaneous velocity is zero and then, as the spring expands, is projected back up the plane, returning to the point at which it was released.

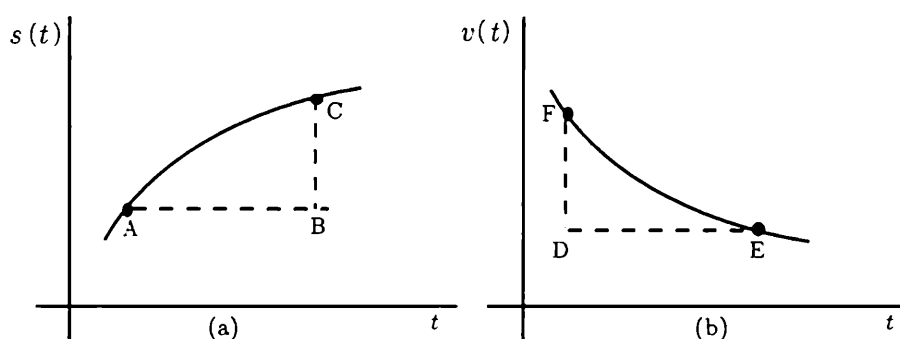
Sketch s versus t and v versus t diagrams for the motion that has been described. Be sure to label the points at which the cart makes contact with the spring, has zero instantaneous velocity, and breaks contact with the spring.

2.31 The diagram shows a velocity vector \vec{v} positioned relative to a set of coordinate axes y and x .

- Construct, on the diagram, the y and x components of the vector \vec{v} . Explain your reasoning by describing how you use the definition of “component” in making your construction.
- Suppose the velocity vector is rotated counter-clockwise through a small angle. Will the two components change in size? If so, describe how each one changes. Does it increase or decrease? Support your answers by making use of the diagram.



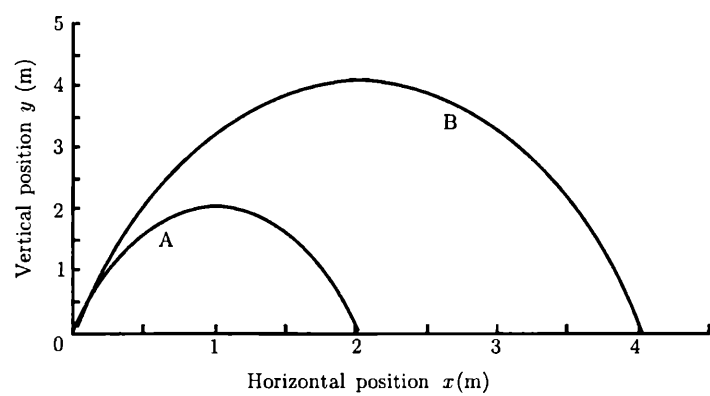
2.32 In these graphs, (a) represents the position versus clock reading history of the rectilinear motion of an object while (b) represents the velocity versus clock reading history of the motion.



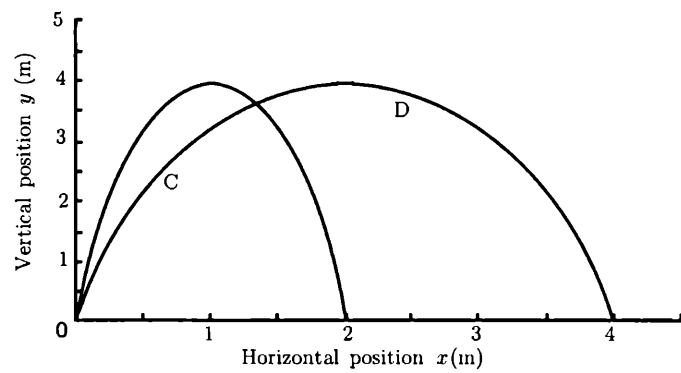
- What is the physical interpretation, if any, of the length of segment AB in diagram (a)?
- What is the physical interpretation, if any, of the length of segment BC in diagram (a)?
- What is the physical interpretation, if any, of the length of the diagonal segment AC (not drawn) in diagram (a)?
- What is the physical interpretation, if any, of the length of segment DE in diagram (b)?
- What is the physical interpretation, if any, of the length of segment DF in diagram (b)?
- What is the physical interpretation, if any, of the length of the diagonal segment FE (not drawn) in diagram (b)?

2.33 The diagrams show the trajectories of two projectiles in the x - y plane under negligible air resistance. Let us denote the magnitude of the initial velocity of the projectile by v_0 and its horizontal and vertical components by v_{0x} and v_{0y} , respectively.

Let us denote the angle of elevation of v_o above the horizontal by θ . When we use the word “compare” in the following questions, we are asking whether one value is greater than, smaller than, or equal to the other. Be sure to explain your reasoning in each instance.



For trajectories A and B: How do the values of v_{ox} compare with each other? the values of v_{oy} ? the values of v_o ? the values of θ ? How do the masses of the projectiles compare with each other?

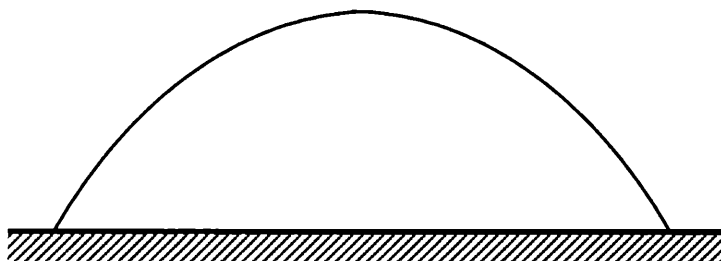


For trajectories C and D: How do the values of v_{ox} compare with each other? the values of v_{oy} ? the values of v_o ? the values of θ ? How do the masses of the projectiles compare with each other?

Chapter 3

Force and Dynamics

3.1 We have established that under the idealization of negligible air resistance, the trajectory of a projectile is a parabola such as that sketched. Let us suppose for the moment that this is the trajectory that would have been followed, in the absence of air resistance, by a ball you have thrown.



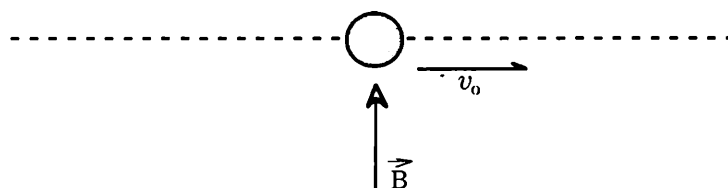
Add to this sketch a trajectory you might expect the ball to follow in the presence of air resistance. You are not being asked for anything quantitative; the idea is to sketch qualitatively the change in shape you might expect. Explain your reasoning.

3.2 It is an observed fact that the force acting on any moving body due to air resistance increases as the velocity of the body increases. Make use of this fact, together with what you know about the motion of falling bodies, to present an argument predicting that raindrops of a fixed size, after accelerating downward for a time interval after their formation, will cease accelerating and attain a constant downward velocity (called “terminal velocity”). In making your argument, be sure to use several force diagrams showing what must be happening to the forces acting on the raindrop during the interval in which it is speeding up.

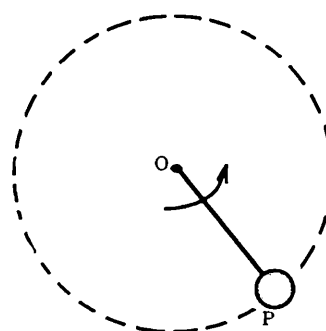
3.3 This figure is a snapshot looking down on a frictionless puck moving at uniform velocity v_o from left to right on a level air table. At the position shown, the puck is given a short, sharp hammer blow \vec{B} in a direction perpendicular to that in which it is initially moving.

- (a) Show on the figure a trajectory (or path) that the puck might follow on the table after the blow is delivered. Explain your reasoning.
- (b) Will the final speed v_f of the puck (immediately after the blow) be equal to, greater than, or smaller than v_o ? Explain your reasoning.

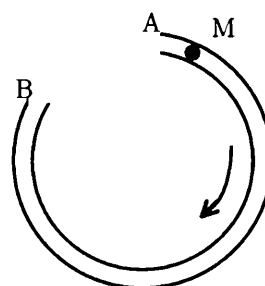
- (c) How will the *velocity* of the puck on the frictionless surface behave as time goes by *after* the blow? That is, will either the magnitude or the direction of the velocity (or both) keep on changing? If so, how?



3.4 In this snapshot, we are looking down on a frictionless puck P which is moving counterclockwise at uniform angular velocity in a circle on a level air table. (The puck is attached to a string, and the end of the string is fastened to a peg at point O.) At the instant the puck is in the position shown, the string is cut. Sketch the trajectory (or path) followed by the puck as it slides along the table after the string is cut. Explain your reasoning. What would have happened if the angular velocity had been increasing at the instant the string was cut? Explain your reasoning.



3.5 A glass tube formed into an incomplete circle lies on a level table as shown. The view is from above. A small marble M is blown into the open end of the tube at A. It swirls around the tube and emerges at the other open end at B. Sketch the trajectory (or path) that the marble will follow as it rolls along the table after leaving the tube at B. Explain your reasoning. How would the path change if the speed of the marble were increased? Decreased? Explain your reasoning.



3.6 Consider a collision between a small car and a heavy truck. In such a collision, how does the size $F_{C \leftarrow T}$ of the force exerted on the car by the truck compare with the size $F_{T \leftarrow C}$ of the force exerted on the truck by the car instant by instant as the two are in contact? Is the first greater than, equal to, or smaller than the second? Explain your reasoning.

3.7 Consider the following statement: "In a tug-of-war, the force exerted by the *losing* side on the winning side must be *greater than* the force exerted by the winning side on the losing side." If you believe the statement to be correct, explain how the winning side manages to win under these circumstances. If you believe the statement to be incorrect, alter it so that it becomes correct and explain your reasoning. Finally, identify the force that makes it possible for the winning side to win.

3.8 Suppose you are pushing a cart along a level floor in the presence of frictional effects between the cart and the floor.

- (a) While you are making the cart speed up, how does the size $F_{C \leftarrow Y}$ of the force you exert on the cart compare with the size $F_{Y \leftarrow C}$ of the force the cart exerts on

you? Is the former greater than, equal to, or smaller than the latter? Explain your reasoning.

- (b) While you are making the cart speed up, how does the size of the frictional force exerted by the floor on the cart compare with the size of the force the cart exerts on you? Explain your reasoning.
- (c) Suppose you are now slowing the cart down from the speed you had imparted to it. How does the size $F_{C \rightarrow Y}$ of the force you now exert on the cart compare with the size $F_{Y \rightarrow C}$ of the force the cart exerts on you? Is the former greater than, equal to, or smaller than the latter? Explain your reasoning.
- (d) While you are slowing the cart down, how might the size of the frictional force exerted by the floor on the cart compare with the size of the force you are exerting on the cart? Explain your reasoning.

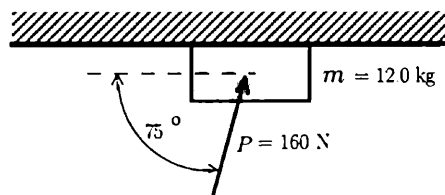
3.9 Suppose you are throwing a ball vertically upward.

- (a) While the ball is still in contact with your hand and you are accelerating it upward, how does the size $F_{B \rightarrow H}$ of the force your hand exerts on the ball compare with the weight W of the ball? Is the former greater than, equal to, or smaller than the latter? Explain your reasoning.
- (b) While the ball is still in contact with your hand and you are accelerating it upward, how does the weight W of the ball compare with the size $F_{H \rightarrow B}$ of the force the ball exerts on your hand? Is the former greater than, equal to, or smaller than the latter? Explain your reasoning.

3.10 Suppose you are accelerating a very massive puck from rest on an air table by exerting a horizontal force. Can you accelerate the puck by exerting a force smaller than the weight of the puck? Explain your reasoning. How small a force will impart at least some acceleration to a very massive puck under perfectly frictionless circumstances? Do you think an ant, harnessed to the puck, would be able to get the puck moving under perfectly frictionless circumstances? Why or why not?

3.11 A block (mass 12.0 kg) is held against the ceiling by a force $P = 160$ N acting at an angle of 75° to the horizontal as shown in the diagram. It is known that the block is in motion (sliding along the ceiling) and that the coefficient of kinetic friction is 0.20.

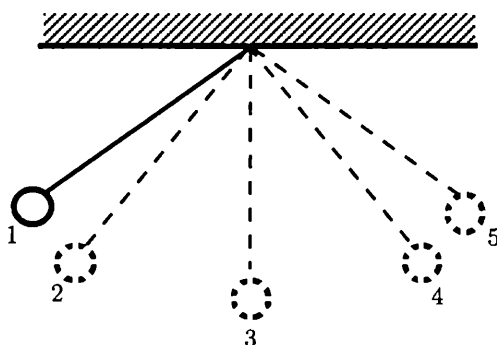
- (a) Calculate the weight of the block. Explain your reasoning briefly.
- (b) Sketch well-separated force diagrams of the block and the region of the ceiling in contact with the block. Describe each force in words and indicate the third law pairs.



- (c) Calculate the normal force exerted by the ceiling on the block by formally applying Newton's second law to the block in the vertical direction.

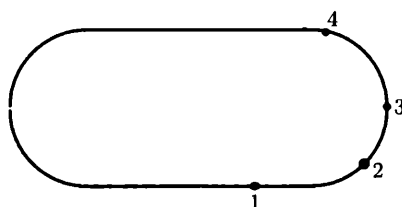
- (d) By formally applying Newton's second law to the block in the horizontal direction, determine whether the block is accelerating along the ceiling, and, if it is, calculate the numerical value of the acceleration. Explain your sequence of reasoning.

3.12 A pendulum bob, let go from rest at position 1, can swing over to position 5 as shown. For *each* of the five indicated positions of the bob, draw three separate vector diagrams for the bob, as described in parts (a), (b), and (c).



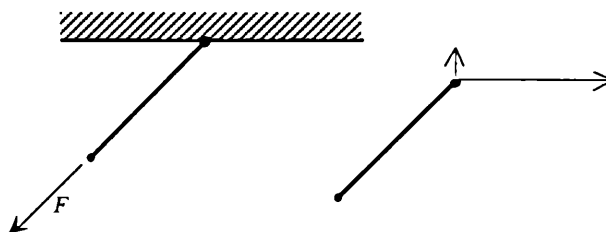
- A diagram showing the instantaneous velocity vector for the bob.
- A diagram showing the forces acting on the bob.
- A diagram showing the instantaneous acceleration vector for the bob. (Keep in mind that the bob is *not* following a rectilinear path.)
- Finally, using vectors from parts (a) and (c), draw vector diagrams showing the *change* in velocity of the bob between positions 4 and 2 and the *change* in acceleration of the bob between positions 4 and 2. Explain how you draw each diagram.

3.13 Here we are looking down on a racetrack with straight sections and semicircular ends. A car is going around the track and maintaining constant speed. For *each one* of the numbered positions draw diagrams showing the velocity vector for the car. On *another* set of diagrams, show the acceleration vector for the car. On a *third* set of diagrams draw the net horizontal force component acting on the car.



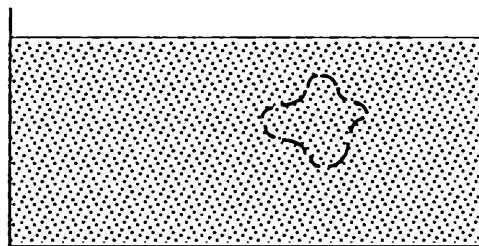
3.14 A string is attached to the ceiling at one end. A person pulls the string at the other end with a force F directed as shown. In the diagram to the right, the person has sketched the horizontal and vertical components of the force exerted by the ceiling on the end of the string.

- (a) Is the right-hand sketch correct and reasonable? Explain your reasoning. (Note that such a sketch may be correct in some respects and incorrect in others.)



- (b) Explain why the string in the diagram is straight. (Hint: Examine the forces acting on chunks of a curvy string, the accelerations that would be imparted to various chunks, and the circumstances under which the accelerations would become zero.) How would the situation change if the direction of the force were reversed? Explain your reasoning.

3.15 Consider a tank or pan containing water, as shown. When the water is first poured into the vessel, it swirls and sloshes around. This motion steadily dies down, however, indicating the existence of a nonzero acceleration and therefore of a net force opposing the motion of every swirling, sloshing parcel of water.



- (a) How do you account for the dying down of the motion? Visualize the forces that might be acting on any arbitrarily shaped parcel of water. What effects can you visualize, even though they are not directly perceptible?

Once the motion in the vessel has died down, the accelerations are everywhere zero, and we infer that every parcel of water must be subjected either to no forces at all or to exactly balanced forces. Consider the particular, irregularly shaped parcel outlined in the figure. This parcel, like any other arbitrarily chosen parcel in the vessel, has its own weight; i.e., the gravitational force exerted by the earth is pulling it downward just as it pulls a book you hold in your hand.

- (b) How do you account for the fact that the outlined parcel of water (and any other parcel of any shape whatsoever) does not fall or accelerate downward? What must be supplying the balancing upward force? Is the effect of the surrounding water distributed or concentrated? If distributed, over what region must it be distributed?
- (c) Draw a force diagram for the outlined parcel, representing the sum of the distributed forces by single arrows (as you would for a book resting on a table) rather than trying to show an actual distribution of forces.

Suppose you take a chunk of wood and submerge it in the water, holding the entire piece below the surface. Note that the wood has displaced (pushed out of the way) a parcel of water of exactly the same size and shape as the wood. As you have indicated in part (c), the original parcel of water that the wood has displaced was being held up by the surrounding water with a total force exactly equal to the weight of the parcel.

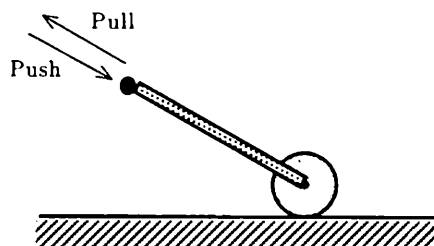
- (d) Is there any reason to believe that this total upward force has changed simply because the original space is now occupied by the new object (the wood) instead of the original water? Explain your reasoning.
- (e) Draw a force diagram for the chunk of wood while you are holding it submerged. Describe each force in words.
- (f) How must the total upward force exerted by the surrounding water on the chunk of wood compare in magnitude with the weight of the chunk of wood: Is it greater, equal, or smaller? Explain your reasoning, noting the role that the relative densities of wood and water play in this context. Explain why you must exert a downward force on the chunk of wood to keep it submerged.
- (g) What will happen to the chunk of wood (in terms of acceleration and change in velocity) after the instant you stop holding it? At what point will the chunk of wood stop moving? Explain in terms of forces acting on the chunk of wood.
- (h) Following a line of argument exactly parallel to that used in connection with the chunk of wood in water, describe how helium or hot air balloons rise in air. Be sure to draw the relevant force diagrams.

Now suppose you take a stone or piece of iron and hold it submerged in water just as you did with the chunk of wood in the preceding analysis.

- (i) Analyze the forces acting on the stone or iron just as you did the forces acting on the chunk of wood in preceding sections. Draw the corresponding force diagrams and answer the corresponding questions. How do you explain the downward acceleration of the stone immediately after you let it go?
- (j) How do you explain the fact that the upward force you exert when you hold the stone submerged in water is perceptibly less than the force you exert when you hold it in air, that is, why does the stone feel “lighter” in water than it does in air?
- (k) How do you explain the fact that a boat made of iron floats even though a simple chunk of iron sinks?
- (l) “Archimedes’s principle” is the name given to the following statement: “An object placed in any fluid is buoyed up by a force equal to the weight of fluid displaced by the object.” What is the justification for this statement? Explain it in terms of the sequence of thinking you have done in this question.

3.16 Consider the situation in which you push or pull a lawn or tennis court roller over the ground as shown.

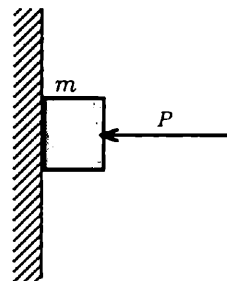
- (a) Draw a force diagram for the roller for the case of pushing, a diagram for the case of pulling, and separate force diagrams for the ground in the region of contact with the roller in each case. Describe each force in words and identify the third law pairs.



- (b) Now examine the following assertion: "It is easier to pull a roller than to push it, but one does a better job of smoothing the ground when one pushes rather than pulls." In the light of the force diagrams you have drawn, do you find this statement to be entirely correct, partly correct, or incorrect? Explain your reasoning.

3.17 A person, exerting a horizontally directed force P , presses a block of mass m against a vertical wall as shown. It is observed that the block does not slide either up or down; it remains at rest in the original position.

- (a) Draw *separate* force diagrams showing (1) the block, with all the forces acting on it, (2) the forces acting on the wall in the region of contact with the block, and (3) the earth and its interaction with the block. Describe each force in words (stating what object exerts this force on what) and identify the third law pairs.

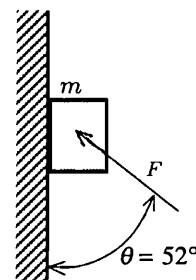


- (b) How do you account, in your force diagrams, for the fact that the block does not slide down along the wall? Describe three separate changes in this situation each of which would lead, without any other changes, to a condition in which the block starts sliding down the wall.
- (c) Suppose m and the coefficient of friction between the wall and the block are known quantities. Describe how you would calculate the value of the force P that is just sufficient to keep the block from sliding. (Your explanation should include an algebraic solution of the problem.)

3.18 A block having a mass of 10.0 kg is pressed against the wall by a hand exerting a force F inclined at an angle θ of 52° to the wall as shown. The coefficient of static friction μ between the block and the wall is 0.20. We shall investigate the question of how large the force F must be to keep the block from sliding along the wall.

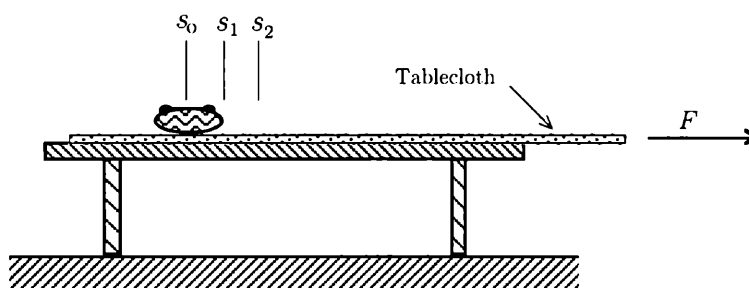
There is more physics here than initially meets the eye. Think about the situation in terms of your everyday experience (or better yet, actually try it out): If you start out with a small value of F , the block will tend to slide downward; as you increase F , you reach the point at which the block will no longer slide; as you continue increasing F , the block stays put until, at some larger value of F , it might even begin to slide upward. This is the physics to be investigated, both algebraically and numerically.

- (a) First draw well-separated force diagrams of the block and the region of the wall where the two are in contact (1) for the case in which F is small enough that the block tends to slide downward and (2) for the case in which the block tends to slide upward. Denote the various forces by appropriate algebraic symbols; do not put in numbers at this point. (The difference between the two sets of diagrams will reside in the direction of the frictional force.) Describe each force in words and identify the third law pairs.



- (b) Applying Newton's second law, obtain algebraic expressions for F in terms of mg , μ , and θ for case 1, in which the block is just about to start sliding downward and for case 2, in which it is just about to start sliding upward.
- (c) Now put in the various numbers and calculate the value of F for each of the two cases. How large is the spread between the two values? Does your result make physical sense? What is going on at the wall when F lies between the two extremes you have calculated? What happens to the frictional force when F lies between these two extremes?
- (d) Return to the algebraic expression for case 2 in which the block is just about to slide upward. What does this expression say happens to F if you keep m and θ constant but increase the value of μ ? What is the equation telling us happens at the point at which μ is large enough to make the denominator of the expression equal to zero? Is it possible to make the block slide upward with a sufficiently large F acting at a fixed value of θ regardless of the value of μ ? Solve for the value of μ at which it becomes impossible to make the block slide upward, showing that this value depends only on θ and is *independent* of the weight of the block. Do you find this result strange? Why or why not? Could you have anticipated it without having made the mathematical analysis?

3.19 Perhaps you have seen the widely performed demonstration in which a tablecloth is quickly yanked out from under a set of dishes on a table; the dishes remain on the table and are not pulled off to fall on the floor. This is usually described as a “demonstration of the effect of inertia.” Let us examine the whole phenomenon carefully.



The figure shows a bowl with mass m_B resting on the tablecloth. A force F is applied to the end of the cloth, and the cloth is yanked out from under the bowl.

The bowl must, of course, be displaced at least some distance along the table. In performing the demonstration, we hope this displacement is small. This is the point we shall examine. Let us subdivide the sliding of the bowl into two obvious stages.

In stage 1 the bowl slides from position s_o (at clock reading t_o) to position s_1 (at clock reading t_1) under the influence of the frictional force exerted by the tablecloth. In stage 2 the bowl slides to a stop at position s_2 (at clock reading t_2) on the table after the cloth is no longer under it. Let us denote the coefficient of sliding friction between the bowl and the cloth by μ_{BC} and the coefficient of sliding friction between the bowl and the table by μ_{BT} .

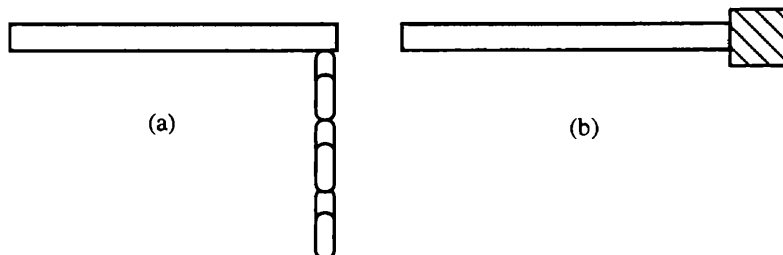
- Draw force diagrams for the bowl in stages 1 and 2, describing each force in words and labeling the various forces with appropriate symbols. Then apply Newton's second law to find the acceleration imparted to the bowl during each stage.
- Now that you have expressions for the acceleration during each stage, use the appropriate kinematic equations to obtain expressions for the two successive displacements $s_1 - s_o$ and $s_2 - s_1$ of the bowl. (Do not lose sight of the fact that the bowl is accelerated to a velocity v_1 at instant t_1 and that it coasts to a stop at instant t_2 . Note that you can now express $t_2 - t_1$ in terms of $t_1 - t_o$.)
- Now show that the total displacement $s_2 - s_o$ of the bowl is given by the following equation:

$$s_2 - s_o = \frac{\mu_{BC} g}{2} \left[1 + \frac{\mu_{BC}}{\mu_{BT}} \right] (t_1 - t_o)^2$$

Note that the time interval appearing in this expression is $t_1 - t_o$, not $t_2 - t_1$.

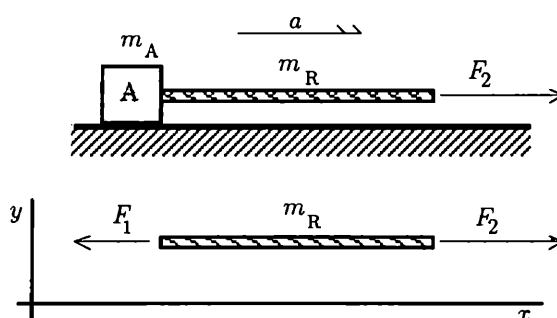
- Interpret the equation: What happens to the displacement $s_2 - s_o$ if μ_{BC} is made very large (i.e., the bowl is virtually glued to the cloth)? What happens to the displacement if μ_{BT} is made very small (i.e., the table is virtually frictionless)? Why is the total time interval $t_2 - t_1$ irrelevant? What happens if the cloth to the left of the bowl (i.e., the length that must slide out from under the bowl) is made longer and longer? Thus, all told, under what circumstances does the demonstration "work" and under what circumstances does it not work?
- Now note that the mass m_B of the bowl does not appear in the final expression for displacement and must therefore be irrelevant to the displacement. In what sense can this experiment be a demonstration of the effect of inertia if the inertial mass of the object subject to the effects does not even appear in the final equation for the displacement???

3.20 Consider the two cases shown, in which a magnet holds various objects while you hold the magnet. In diagram (a) the magnet holds a string of iron paper clips hanging end to end; in diagram (b) it holds a small block of iron.



- (a) Draw separate force diagrams in case (a) for each paper clip and for the magnet. Describe each force in words and identify the third law pairs. (Show larger forces with longer arrows and equal forces with arrows of equal length.) How do you account for that fact that the paper clips that are not in direct contact with the magnet do not fall?
- (b) Draw separate force diagrams in case (b) for the iron block and the magnet. Describe each force in words and identify the third law pairs. How do you account for the fact that the iron block does not fall?

3.21 Suppose we apply the force F_2 to the end of a rope having mass m_R and proceed to accelerate the system, consisting of the massive rope and block A with mass m_A , in the positive x direction along a surface with negligible friction.



A force diagram for the rope is shown, with the force F_1 drawn smaller than F_2 because the rope is being accelerated and must therefore have a net force acting on it in the horizontal direction. The acceleration of the entire system is denoted by a , since we assume that the rope does not stretch. (What could we say about the acceleration if the rope were stretching?) In the following, we shall concern ourselves with the horizontal forces only and treat vertical forces (e.g., weight of the rope) as unimportant to the physics under consideration.

- (a) Draw the force diagram for block A and argue (1) that $F_1 = m_A a$, and (2) that $F_2 - F_1 = m_R a$. How does the magnitude of F_1 compare with the magnitude of F_2 ? How do you explain the difference physically?

We apply the adjective “tensile” to forces such as F_1 and F_2 acting on a rope, string, rod, bar, or any object being stretched as the rope, in this example, is being stretched. We say that F_1 is the “tension” at the plane cross section through the left-hand end of the rope and that F_2 is the “tension” at the plane cross section through the right-hand end. Similarly, we call the force acting on a plane cross section through *any* location along the rope “the tension at that location.” In the light of this definition, we see that the force to which we have given the name “tension in the rope” is not uniform but varies continuously from one end to the other when the system is being accelerated.

- (b) Consider the right-hand half of the rope as a chunk with horizontal forces acting on it. Draw a force diagram of this right-hand half, denoting the force at the left-hand end of this chunk by F_x . Draw a force diagram for the rest of the rope.

In the light of our definition, what is an appropriate symbol for the tension in the rope at the left-hand end of the right-hand half? How would you expect this tension to compare in magnitude with F_2 and F_1 ?

- (c) Let us examine the *difference* between F_2 and F_1 as compared with the magnitude of F_2 . That is, make use of the results in part (a) to show that

$$\frac{F_2 - F_1}{F_2} = \frac{m_R}{m_R + m_A}$$

- (d) Analyze and interpret this equation: (1) How do the tensions at the two ends of the rope compare (for a fixed acceleration a) if m_R is made smaller and smaller relative to m_A ? (2) If m_A becomes indefinitely large relative to m_R ? (3) What happens to F_1 as m_A is made smaller and smaller relative to m_R ? (4) What is the tension at the left-hand end of the rope if it is being accelerated with the left end free, i.e., with nothing attached to it?
- (e) Under what circumstances is tension completely uniform throughout the entire length of a rope? Under what circumstances of m_R compared with m_A may the tension be regarded as *very nearly* uniform while the system is accelerating?
- (f) In the light of your answers to the preceding questions, what is the real meaning of the term “massless string” as it is used in many physics problems you have encountered?

Now consider some situations in which a line of discrete objects (that are connected to each other) is being accelerated by a force applied at one end as follows:



This might be a group of blocks connected by “massless” strings as implied in the diagram. It might be a long chain. It might be a string of freight cars on a railroad track.

- (g) Discuss these seemingly different situations qualitatively. In what ways are they basically similar? How do the forces acting on each successive object compare with the forces on the object to the right? How does the force to the right on each successive object compare with the magnitude of the force F ?
- (h) Derive an expression for the force to the right on the n th object from the right, assuming the masses of the objects to be identical.
- (i) Suppose the line of objects in the preceding figure is connected by identical springs that stretch somewhat in accordance with Hooke’s law. If the objects are equally spaced before the system is accelerated (as illustrated), what will the spacing be like after all the objects in the system have acquired the same acceleration?

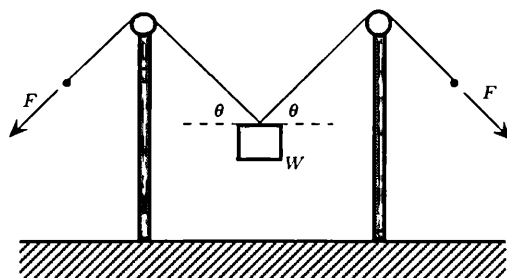
- (j) If, in the situation under consideration in part (i), the force is applied to the lead object abruptly at some instant, will all the objects be accelerated at that same instant? Describe what will actually happen in such a system under these circumstances. You are not being asked to calculate or derive anything; just visualize physically what will happen between the initial instant at which the force is applied and the instant at which all the objects finally have the same acceleration. Under what circumstances will the time interval involved be very short? Under what circumstances might it become quite long?
- (k) In the light of your discussion in parts (i) and (j), what do you *visualize* happens in the massive rope after you apply an accelerating force abruptly at one end, even though you cannot actually *see* what happens?

3.22 A heavy rope with total mass m hangs from the ceiling. What is the tension in the rope

- at the section at which the rope is fastened to the ceiling?
- at the bottom of the rope?
- at the middle of the rope?

3.23 Any elastic cord, such as a bungee jumping cord, has an effective spring constant like that of any spring. Will a shorter bungee jumping cord have a larger or a smaller spring constant than a longer cord? Explain your reasoning.

3.24 Suppose we have an object of weight W suspended on a string between poles as shown. We proceed to elevate the object by pulling with forces F on the ends of the string. As we do so, the angle θ decreases, and the string becomes more nearly horizontal.

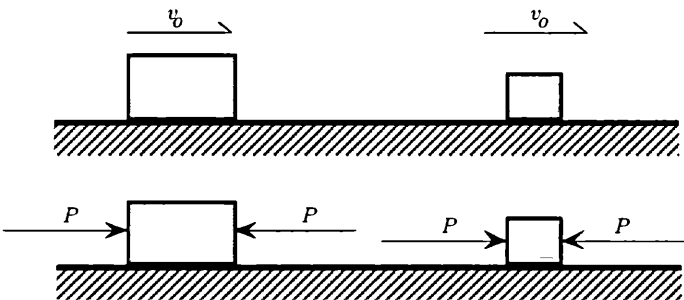


- Draw the relevant force diagram, and show that $F = (W/2) \sin \theta$ if we neglect the weight of the string.
- Interpret the equation in part (a). What is the value of F when θ is close to 90° ? What happens to the magnitude of F as the string becomes more and more nearly horizontal? What is the possibility of getting the string to be absolutely horizontal as long as W is not equal to zero?
- In the light of the analysis in part (b), and recognizing that no string can be completely massless, comment on the following little rhyme:

“No force, however great,
Can stretch a thread, however fine,
Into a horizontal line
That shall be absolutely straight.”

Is this nonsense or is it physically correct? Explain your reasoning

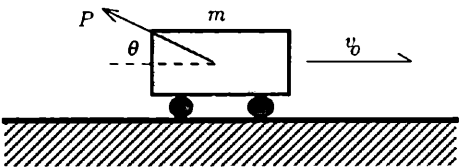
3.25 Two blocks of dry ice with obviously different masses are sliding at the same velocity v_0 on a glass plate and staying a fixed distance apart, as shown in the upper part of the diagram. The motion is very nearly frictionless since the blocks are sliding on the layer of carbon dioxide gas between their bottom surfaces and the surface of the glass plate. Forces P of identical magnitude are suddenly and simultaneously applied to both blocks as shown in the lower part of the figure.



Describe what happens to the motion of each block. Do the blocks continue to stay the same distance apart? If their motions are different after application of the forces, how do they differ? Explain your reasoning.

3.26 A cart with mass m moves to the right on a horizontal surface under conditions of negligible frictional resistance. At instant $t = 0$ the velocity of the cart is v_0 , and its position is at $x = 0$. At $t = 0$, the force P is suddenly applied at an angle θ as shown. The force remains constant in magnitude after it is applied.

- (a) Draw separate force diagrams of the cart and the region of the surface in contact with the cart.
- (b) Applying Newton’s second law, solve algebraically for the force exerted by the surface on the cart. (Be sure to indicate your choice of positive direction.)
- (c) Applying Newton’s second law, solve algebraically for the acceleration imparted to the cart.

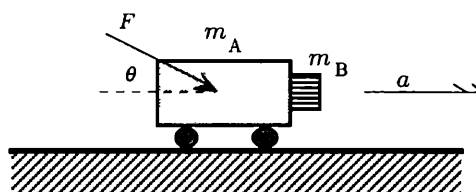


Suppose the cart has a mass of 15.0 kg, the initial velocity to the right is 5.8 m/s, the magnitude of the force P is 18.0 N, and the angle θ is 27° .

- (d) Using the algebraic result you obtained above, calculate the magnitude of the normal force exerted by the surface on the cart, being careful to give only the proper number of significant figures in the final result. Interpret the result: How does the normal force compare in magnitude with the weight of the cart? Does your result make physical sense? Why or why not?

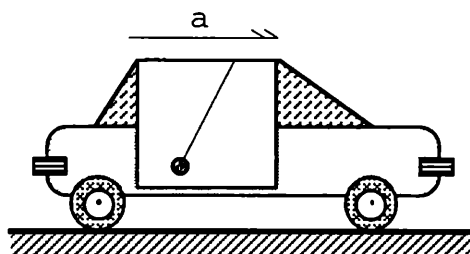
- (e) Using the procedures developed in earlier study of kinematics, calculate where the cart will be located 12 s after the force P is applied. Interpret the result; i.e., describe the motion and successive positions of the cart between clock readings $t = 0$ and $t = 12$ s.
- (f) Suppose the cart had a mass of 0.50 kg instead of 15.0 kg. What would happen on application of force P of 18.0 N? Explain your reasoning.

3.27 A force F , acting as shown, imparts an acceleration a to the system consisting of cart A and block B. Block B simply rests against the right-hand wall of the cart; it is not fastened to the cart. The coefficient of static friction between the surfaces of A and B is denoted by μ



- (a) Draw separate force diagrams for bodies A and B and describe each force in words.
- (b) The problem to be investigated is that of how large the acceleration a must be to keep block B from sliding downward. Analyze the situation and comment on the feasibility of achieving the condition indicated, assuming realistic values of forces, masses, and acceleration. (Note: Not all the parameters indicated in the diagram are necessarily relevant to the analysis.)

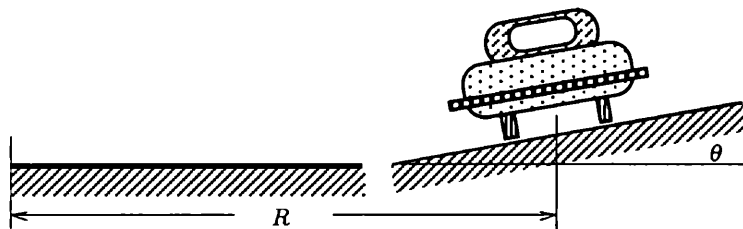
3.28 A pendulum hangs from the roof of a car that is accelerating to the right as shown. Under these circumstances, the pendulum hangs as shown.



- (a) Draw well-separated force diagrams for the bob, the string, and the region of the car roof where the string is attached. Describe each force in words and identify the third law pairs.
- (b) Show how the pendulum would hang if the car were moving to the right at high constant velocity. Explain your reasoning.
- (c) Show how the pendulum would hang if the car were slowing down while moving to the right. Explain your reasoning.
- (d) Show how the pendulum would hang if the car were slowing down while moving to the left.

3.29 A car is traveling away from us on a banked road having a curve of radius R . The car is traveling at a speed greater than that for which the banking angle θ has the optimum value.

- (a) Draw well-separated force diagrams for the car and for the road surface at the location of the car. Describe each force in words and identify the third law pairs. (Confine yourself to forces that lie in the plane shown in the diagram; do not try to include forces that act into or out of this plane.)
- (b) Draw the same force diagrams as in part (a) for the case in which the speed of the car is less than that for which θ is the optimum angle of banking.
- (c) Draw the same force diagrams as in part (a) for the case in which the speed of the car is equal to that for which θ is the optimum angle of banking.

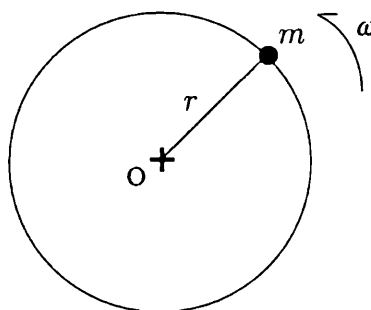


- (d) Describe in your own words the physical differences among the three cases you have illustrated. Pay particular attention to the role of frictional effects parallel to the road surface.

3.30 Consider the situation in which a bob on a string is caused to revolve in a circle of radius r around a fixed point O as shown. The bob has mass m and instantaneous angular velocity ω . The motion takes place in a *vertical* plane.

- (a) First consider the situation at the very top of the circle. Draw force diagrams for both the bob and the string at that instant. Apply Newton's second law to the motion of the bob and show that if the force exerted by the string on the bob is denoted by T and positive direction is taken in the direction of the centripetal acceleration,

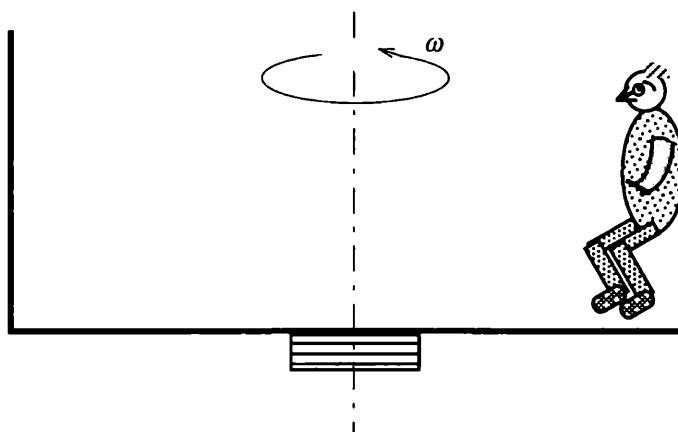
$$T = mr\omega^2 - mg$$



- (b) Interpret this relation by examining what it tells us happens to T as we imagine starting with a large value of ω and decrease ω continuously. What is happening when $T = 0$? How does the bob behave when ω is smaller than the value that makes $T = 0$? How might you interpret the negative values of T that are indicated by the equation when ω becomes sufficiently small? What will happen if we make ω larger and larger without limit?
- (c) Examine the situation and behavior of the bob at the bottom of the circle following a sequence parallel to that outlined in parts (a) and (b).
- (d) Now, in a parallel analysis, examine what happens to a roller coaster car as it goes over the top of a circular track and as it passes through the bottom of a circular track. What force replaces the effect of the string in the preceding analysis? What happens at the top of a circle if the car is going too rapidly?

- (e) Apply a parallel analysis to what happens to an object going around a vertical loop-the-loop. How does the situation at the top of the loop-the-loop differ from that at the top of the roller coaster? In what way is the situation at the top of the loop-the-loop similar to that of the bob on the string? In what way is it different?

3.31 A cylindrical chamber in an amusement park rotates around a vertical axis as shown in the following diagram. When the angular velocity is sufficiently high, a person leaning against the wall can take his or her feet off the floor and remain “stuck” to the wall without falling.



- (a) For these circumstances, draw force diagrams of the person and for the region of the wall in contact with the person. Describe each force in words and indicate the third law pairs.
- (b) Take the coefficient of friction between the person and the wall to be 1.2. Assume a reasonable mass for the person. Investigate what the period of rotation of the chamber would have to be for the person to be able to take feet off the floor at various chamber radii without falling. Select a combination of values of period and radius that you would consider reasonable and feasible and justify your selection. Explain your reasoning throughout. Would it be easier to produce the desired effect if the coefficient of friction were larger? How large would you like it to be?
- (c) Suppose the person were to put a basketball against the wall at the same height as his head while the system is rotating and the feet are off the floor. Draw a force diagram for the ball and describe each force in words. How will the ball behave? Will it stay where it is placed or will something else happen? Explain your reasoning.

3.32 A person, with mass m_p , stands at the rim of a merry-go-round holding a pendulum bob, of mass m_b , on a string. The merry-go-round has a radius R and rotates at a constant angular velocity ω . As the rider adopts the most comfortable stance under the circumstances, a photograph shows him or her to be leaning in such a way that the body line is exactly parallel to the string of the pendulum.

- (a) Draw force diagrams for the person, the pendulum, and the region of the merry-go-round in contact with the person. Describe each force in words and identify the third law pairs.
- (b) In terms of the ideas of force and acceleration we have been studying, explain why the person adopts the body angle observed. In what sense is it “most comfortable”? In the final analysis, are any of the parameters mentioned above irrelevant to the explanation you have given? If so, which ones? Explain your reasoning.
- (c) Suppose you wish to calculate the actual angle relative to the vertical adopted by the pendulum string in particular circumstances. Which parameters are needed for the calculation and which, if any, are irrelevant? Explain your reasoning.

3.33 Consider a pendulum bob suspended freely from the ceiling or some other support at each of the following locations at the surface of the earth: The north pole, the equator, and some intermediate latitude. Let us take the earth to be perfectly spherical (we know this is not actually the case) even though it is rotating. Sketch how the string on which the bob hangs would be oriented relative to a radial line from the center of the spherical earth at each of the three locations (a) if the earth were not rotating and (b) with the earth rotating.

3.34 Be able to explain and interpret the following experiment described by Newton in the *Principia* (words or phrases in brackets [] are our editorial insertions to assist the modern reader):

“I tried the thing in gold, silver, lead, glass, sand, common salt, wood, water, and wheat. I provided two equal wooden boxes. I filled one with wood, and I suspended an equal weight of gold (exactly as I could) in the center of oscillation of the other. The boxes, hung by equal threads of 11 feet, made a couple of pendulums perfectly equal in weight and figure, and equally exposed to the resistance of the air. Placing the one by the other, I observed them to [swing] together forwards and backwards for a long while with equal vibrations. And therefore the quantity of matter [inertial mass] in the gold was to the quantity of matter in the wood as the action of the motive force [gravitational force] upon all the gold to the action of the same upon all the wood; that is, the weight of one to the weight of the other.”

What was the point of this experiment? What did Newton observe? What did he infer from the results?

3.35 Explain the basis for acceptance of the model according to which, in the solar system, the earth and planets all revolve around the sun rather than the model in which all the other members revolve around the earth. In other words, in what sense do we come to accept the heliocentric solar system rather than the geocentric one? Keep in mind the fact that we do not completely reject and abandon the geocentric model; we use it continually, for example, when we navigate on the surface of the earth. [This is not a simple question with a short, pat answer. It involves a lengthy story of successes stemming from a physical theory. We do not establish the model through direct observation.]

3.36 Suppose that, in a hypothetical other world, the law of gravitation takes the form

$$F_{\text{grav}} = K \frac{\sqrt{m_1 m_2}}{r^2}$$

while Newton's laws of motion are otherwise valid, i.e., $F_{\text{net}} = ma$. (The symbol m denotes inertial mass.)

- (a) Suppose m_2 represents the mass of a planet and m_1 that of a falling body. Examine the algebraic implications: How would freely falling bodies behave on a planet in this world? Would objects of smaller mass fall with the same acceleration, smaller acceleration, or greater acceleration than bodies of larger mass? Explain your reasoning.
- (b) Give a numerical example of the conclusion reached in part (a), e.g., If an object with a mass of 1.0 kg falls with an acceleration of 10 m/(s)(s), what would be the free fall acceleration of a rain drop with a mass of 0.1 g? In the absence of air resistance, what would be the free fall velocity at the end of one second of each of the two objects (falling from rest)? How would you like to get hit by such a rain drop?
- (c) Invent a gravitation law in which free fall acceleration would *increase* with inertial mass.

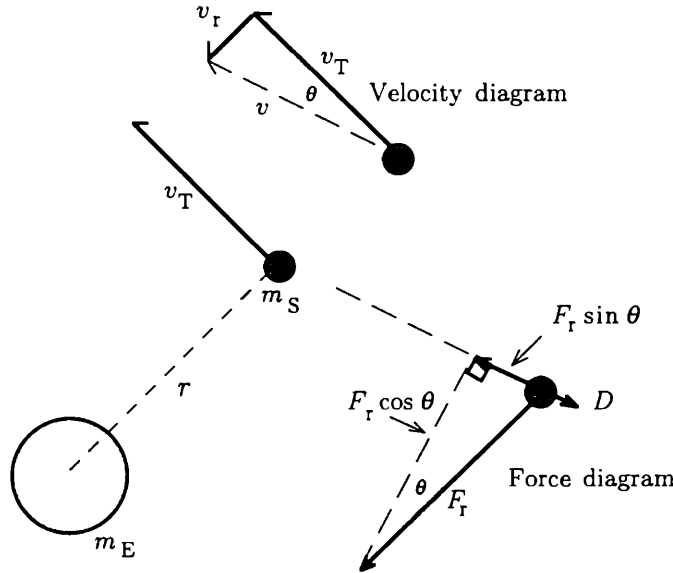
3.37 Consider the case of a satellite of mass m_s in circular earth orbit. Any satellite orbiting in the upper reaches of the atmosphere is subject to a drag force D through collision with molecules of gas that are present in the tenuous upper atmosphere. Under the influence of the drag force, the satellite tends to spiral slowly in toward smaller orbital radii. In the process, the orbital velocity of the satellite *increases*. How can a body speed up under the influence of a retarding (drag) force? What is the effective force that must be accelerating the satellite under these circumstances? We shall analyze this situation in detail and arrive at the surprising conclusion that the force accelerating the satellite is actually equal in magnitude to the drag force D ! The overall effect is as though we turned D around and made it accelerate the object whose motion it is retarding.

The situation we are analyzing is sketched in the following diagram. The diagram also shows separate velocity and force diagrams for the satellite. The mass of the earth is denoted by m_E . The tangential velocity of the satellite at orbital radius r is denoted by v_T . The gravitational force exerted by the earth on the satellite is denoted by F_r .

Let us first note, qualitatively, the effect of the drag force as sketched in the velocity diagram. The drag force D , opposing the motion, causes the satellite to spiral inward. Thus the satellite acquires a very small radial velocity of magnitude v_r , which is greatly exaggerated in the velocity diagram to make the effect visible. With this very small radial velocity, the satellite acquires a total vector velocity of magnitude v and follows a descending path, the tangent to which makes the small angle θ with the direction of the tangential velocity (magnitude v_T) in the circular orbit that would have been followed in the absence of drag. We shall call θ the "angle of descent." The drag force D lies along the line of descent, in the direction opposite to v , as shown in the force diagram. Since v_r is vastly smaller than the tangential velocity v_T , we shall keep using v_T as an adequate approximation to the magnitude

of the total velocity v . Thus the application of the drag force causes the satellite to settle down into an angle of descent θ with a radially inward velocity v_r . Since v_r , θ , and the acceleration a_D along the line of descent all depend on D , our problem is to solve for these three relationships, i.e., for each of the three quantities in terms of D . It emerges that $m_S a_D = D$!

Now let us turn to the force diagram. The total force acting on the satellite is actually the resultant of the drag force D and the centripetal gravitational force F_r . Since D , however, is vastly smaller than F_r , we can see what is happening without trying to construct the resultant but just by looking at the effect of F_r along the line of descent of the satellite. Note that F_r has a nonzero component $F_r \sin \theta$ along the line of descent and in the direction of motion of the satellite. The resultant acceleration a_D imparted to the satellite along the line of descent must be imparted by the force $F_r \sin \theta - D$ acting along this line. In our analysis of what is happening we must study this force.



- (a) First show how we arrive at the following two basic equations and describe the physical meaning of the quantity $m_S a_D$ in your own words:

$$F_r \sin \theta - D = m_S a_D \quad (1)$$

$$F_r = \frac{m_S v_T^2}{r} = G \frac{m_S m_E}{r^2} \quad (2)$$

- (b) We can find out more about a_D and its relation to the various forces and velocities by powerful use of the chain rule for differentiation:

$$a_D = \frac{dv}{dt} = \frac{dv}{dr} \frac{dr}{dt} = v_r \frac{dv}{dr} \cong v_r \frac{dv_T}{dr} \quad (3)$$

Explain Eq. 3 in your own words. Where did v_r come from? Why the \cong sign?

- (c) We can now make use of Eqs. 2 and 3 to show that:

$$m_S a_D = \frac{1}{2} F_r \sin \theta \quad (4)$$

Fill in the steps leading to Eq. 4. [Hint: Differentiate Eq. 2 to find dv_T/dr and use the velocity diagram to show that $v_r \cong -v_T \sin \theta$. (The negative sign is necessary because v_r is negatively directed.)]

- (d) By eliminating $F_r \sin \theta$ from Eqs. 4 and 1, show that

$$m_S a_D = D \quad (5)$$

and also show that

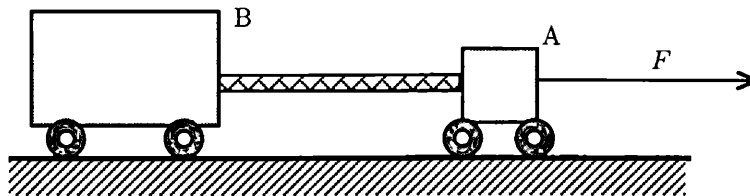
$$F_r \sin \theta = 2D \quad (6)$$

$$\sin \theta = \frac{2D}{F_r} \quad (7)$$

$$v_r = -2D \sqrt{\frac{r}{m_S F_r}} \quad (8)$$

- (e) Interpret the results obtained in part (d) both in your own words and by sketching a force diagram that shows the relationships. Check whether or not the results are dimensionally consistent. What is the point of obtaining Eqs. 7 and 8 in addition to 5? What is the role of the gravitational force (in addition to that of the drag force) in determining the direction of tangential acceleration? What happens to both v_r and θ as the drag force is increased? Does this make physical sense? Why or why not?

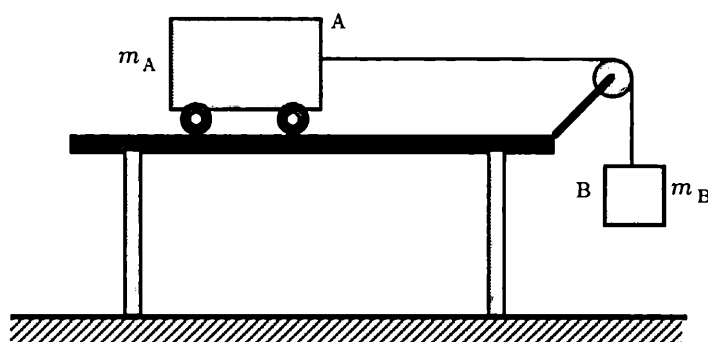
3.38 Suppose two carts, connected by a rope, are being accelerated by a force F as shown. The carts have very different masses, but the frictional forces acting on them are nearly the same. The rope is, of course, under tension in these circumstances. Suppose the force F remains unchanged, but the order of the carts is reversed, i.e., B is placed on the right and A on the left. How does the tension at the center of the rope now compare with what it was before the reversal of the carts? That is, is the tension larger than, less than, or the same as it was initially? Explain your reasoning.



3.39 The law of universal gravitation is given by the equation $F_{\text{grav}} = Gm_1m_2/r^2$, but the expression for the gravitational force exerted by the earth on an object near its surface (the “weight” of the object) is given by mg . What has happened to r ? Why is r absent from the latter expression?

3.40 Consider the situation illustrated below: The system accelerates when object B is released. The inertial effects of both the string and the pulley are negligible. The string is essentially unstretchable.

- (a) How does the acceleration of body A compare with that of body B? Are the accelerations equal or different? If different, which is larger? Explain your reasoning.
- (b) While the system is accelerating, how does the magnitude of the force exerted by the string on cart A compare with the weight of body B? Is the force exerted by the string on cart A equal to, greater than, or smaller than the weight of B? (An algebraic solution is not being called for; the reasoning should be performed qualitatively.) Explain your reasoning.



- (c) Suppose that m_B is increased while m_A remains unchanged. What will happen to the acceleration of the system? Will it increase, decrease, or remain unchanged? What will happen to the force the string exerts on body A? What will happen as m_B is increased further and m_A becomes very small relative to m_B ? Explain your reasoning.
- (d) Describe what will happen to the acceleration of the system and to the force the string exerts on body A if the changes in part (c) are reversed and m_B becomes very small relative to m_A . Explain your reasoning.
- (e) Suppose carts A and B are connected by a rope with significant mass rather than by a massless string. Will the acceleration of the system be uniform or nonuniform? If nonuniform, will the acceleration increase or decrease after release of cart B? Explain your reasoning.
- (f) What difficulties would be introduced into the problem if the string or rope connecting the two bodies were stretchable rather than unstretchable?

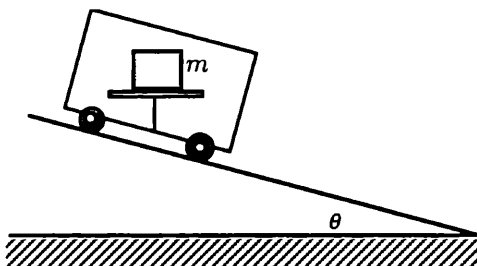
3.41 Suppose you were to set up a seesaw in an elevator that could be accelerated either up or down.

- (a) Imagine first an experiment in which you have a partner of mass equal to your own, and you balance the seesaw while the elevator is at rest. Now, with the seesaw initially balanced, the elevator is accelerated upward. Will the seesaw remain balanced during upward or downward acceleration? Explain your reasoning, being sure to draw separate force diagrams for yourself, your partner, and the seesaw.

- (b) Now imagine a second experiment in which your partner's mass is smaller than yours. You conduct the same experiment in the accelerating elevator starting with a balanced seesaw. Will the seesaw remain balanced during upward or downward acceleration? Explain your reasoning in a manner similar to that in part (a), being sure to draw the relevant diagrams.

3.42 Suppose a cart is accelerating freely down an inclined plane as shown. Inside the cart is a horizontally fixed platform on which rests a block of mass m .

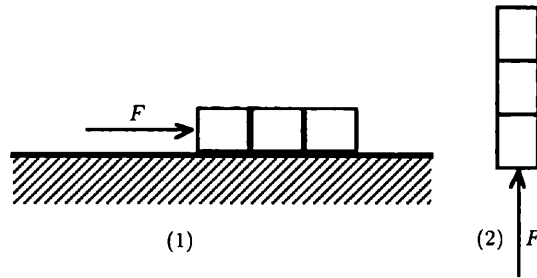
- (a) Suppose the platform is frictionless. Will the block appear to stay in a fixed position within the cart or will it slide relative to the platform? If it shifts its position within the cart, which way does it shift? If the block tends to slide on the frictionless platform, in what direction must a frictional force act to keep it from sliding? Explain your reasoning.



- (b) Suppose the block is resting on a platform balance while the system is accelerating. (There is sufficient friction between the block and the surface of the platform balance to keep the block from shifting horizontally relative to the cart.) How will the reading on the balance compare with the weight of the block? That is, will it be equal to, greater than, or smaller than the weight of the block? Explain your reasoning.
- (c) How will the reading on the balance in part (b) compare with the magnitude of the normal force that would be exerted by the plane on the block if the block were simply sliding down the plane in the absence of friction?
- (d) Analyze the situation in part (b) algebraically, and show that the horizontal frictional force must be equal to $mg \sin \theta \cos \theta$ and that the reading on the platform balance must be equal to $mg \cos^2 \theta$.

3.43 A block rests on a level floor, and an external horizontal force F is exerted on it. There is friction between the block and the floor. The force F starts at zero and is increased slowly in magnitude until the block begins to slide. Describe in your own words the behavior of the frictional force during the time interval between the instant of initial application of F and the instant at which the block begins to slide. Is the frictional force zero until the block begins to slide? Is it constant? Does it vary over this time interval? If so, how does it vary? Explain your reasoning.

3.44 In case (1) a group of three identical blocks in contact with each other is being accelerated along a frictionless horizontal plane by the constant force F . In case (2) the same group of three blocks is being accelerated vertically upward by the same constant force F .



- (a) Draw a separate, complete force diagram for each of the three blocks in each of the two cases, comparing the forces on each block, and from block to block, by using longer arrows for larger forces, shorter arrows for smaller forces, and arrows of equal length for equal forces. Draw a separate force diagram for the horizontal plane in case (1), showing the forces of interaction between the blocks and the plane. Describe each force in words by indicating the nature of the force (gravitational, mechanical contact, etc.) and stating what object exerts the given force on what.
- (b) Referring to your diagrams, describe in your own words the similarities between cases (1) and (2) and the essential differences.
- (c) Suppose that in case (2) a cylinder of water that is being accelerated vertically upward instead of three separate blocks. What similarities do you see between the case of the cylinder of water and the case of the three blocks? What differences?
- (d) What happens if case (1) is a cylinder of water instead of three separate blocks?

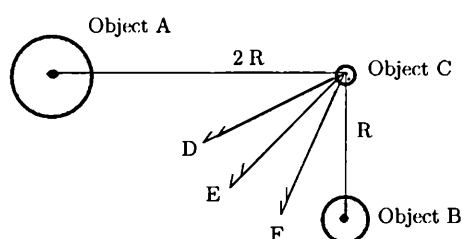
3.45 Along with all the other objects at rest at the surface of the earth, you make a complete rotation around the earth's axis in 24 hours. Is it legitimate to say that you and the other objects are "in orbit" around the earth's axis? Why or why not? (In the course of your explanation, be sure to examine what "being in orbit" means.)

3.46 Can a satellite, which is to be propelled into earth orbit, be put into an orbit that lies in the same plane as the one in which you rotate around the earth's axis in the position at which you are located on the surface of the earth? Why or why not? Sketch a relevant diagram to assist your explanation.

3.47 Which of the following facts provide supporting evidence for Newton's hypothesis that the force of gravity is proportional to the inertial masses of the interacting bodies?

- (a) Kepler's laws are obeyed by planets of very different masses.
- (b) The period of an earth satellite is independent of its mass.
- (c) Falling bodies all have the same acceleration when air resistance is negligible.
- (d) Each of the foregoing facts provides some evidence.
- (e) None of the facts (a), (b), or (c) provides evidence for this aspect of the interaction.

3.48 Here object A has four times the mass of object B. The objects A and B are fixed in space and cannot move. The small object C is instantaneously located, relative to A and B at the position shown. Which arrow in the diagram best shows the direction in which C would be accelerated by the gravitational forces exerted by A and B at the instant under consideration. Explain your reasoning.



- (a) Arrow D.
- (b) Arrow E.
- (c) Arrow F.
- (d) Arrow D, E, or F, depending on the direction and magnitude of the instantaneous velocity of C.
- (e) Arrow D, E, or F, depending on whether A and/or B are also free to move.

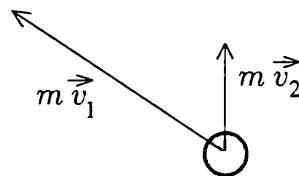
3.49 Consider each of the following statements. If you believe a statement is consistent with Newton's laws of motion, mark it with Y for "yes" and give a specific example of a situation of the kind described. If you believe a statement is not consistent with Newton's laws, mark it with an N for "no" and explain what is wrong with it.

- (a) A body exerts two different forces on another object.
- (b) The earth exerts a force on an object in outer space and the object exerts an equal and opposite force on the earth.
- (c) A body moves at uniform velocity with only one force acting on it.
- (d) A body in outer space accelerates under the influence of a force but exerts no forces on any other object.
- (e) During the interval of collision between a small car and a large truck, the small car is subjected to a larger force than is the large truck. This difference accounts for the greater damage sustained by the small car.
- (f) Applying the brakes cannot stop a car because the brakes exert a force internal to the car and internal forces cannot accelerate an object.

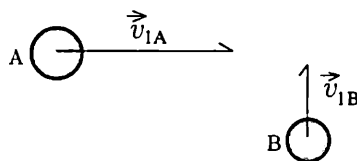
Chapter 4

Momentum and Energy

4.1 A particle of mass m has an initial momentum vector $m\vec{v}_1$ as shown. After being given a sharp blow, the particle has a final momentum vector $m\vec{v}_2$. Use the two arrows to construct a vector representing the impulse \vec{I} that must have been delivered to the particle by the blow that was imparted. Explain your reasoning.



4.2 Two frictionless pucks A and B moving on an air table have velocities \vec{v}_{1A} and \vec{v}_{1B} , respectively. The pucks collide and stick together in a perfectly inelastic collision. Puck A has just twice the mass of puck B.

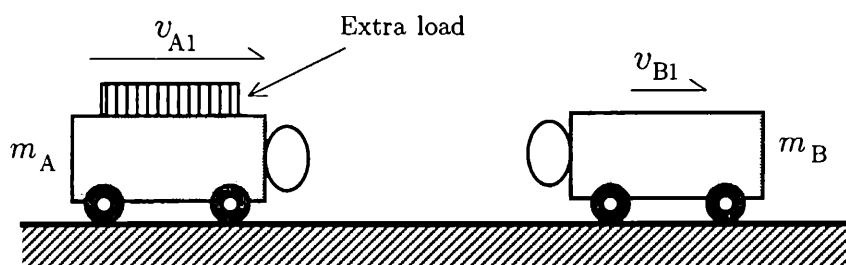


- Construct two vectors representing the momenta of the two bodies just before collision.
- Making use of the vectors in part (a), construct a vector that shows the momentum of the system of the two pucks just after the collision. Explain your reasoning.
- What impulse was imparted to the system (consisting of the two pucks) between the initial and final conditions? What impulse was imparted to each individual puck? Construct the relevant vectors and explain your reasoning.

4.3 Consider two carts with masses m_A and m_B , respectively, equipped with very soft spring bumpers that undergo quite large displacements as they are compressed in a rectilinear collision. (A similar situation can be set up using gliders on an air track.) When the extra load is removed from cart A, the masses of the two carts are identical. The velocity magnitudes of the carts or gliders before collision are denoted by v_{A1} and v_{B1} , respectively. The purpose of the very soft spring bumpers is to allow one to see the actual compression and springing apart as collision occurs. In performing the following experiments, make the collisions fairly gentle so as not to ruin the bumpers.

- Perform some *qualitative* experiments with a system of this kind, starting with body B stationary ($v_{B1} = 0$), and observing what happens under various circumstances without making numerical measurements: (1) What happens when

neither cart is loaded and $m_A = m_B$? (2) What happens when cart A is loaded so that $m_A > m_B$? (3) What happens when cart B is loaded so that $m_A < m_B$?



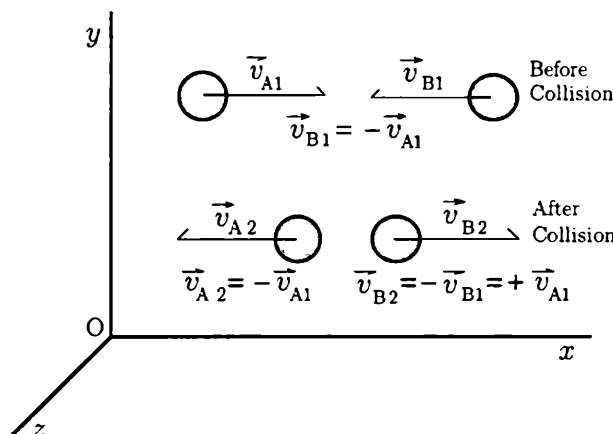
- (b) What happens when the carts approach each other with approximately equal but opposite velocities?
- (c) In at least a few of the experiments, sketch what the graph of force versus clock reading must look like for each cart over the interval from just before to just after contact. Directly underneath this graph and using the same scale for clock readings, sketch what the acceleration and velocity graphs must look like. Then sketch what the corresponding graphs for collisions between billiard balls or steel ball bearings (where it is impossible to see the deformations that take place) might look like on a similar time scale. What determines the time interval of interaction in these widely different circumstances?
- (d) Now consider the collision between carts or gliders equipped with strong magnets mounted so as to repel each other. Carry out observations of collisions, such as those in part (a), if at all possible. Make the collisions quite gentle so that the magnets do not actually strike each other but the carts still spring apart. (Even though the carts never touch each other, we still call this a collision.) What is happening in the absence of physical contact? What do the various graphs look like in such cases? What is the time interval of interaction, and how is this represented on the graphs?
- (e) Suppose the magnets were mounted so as to attract rather than repel each other (or imagine the colliding objects to be carrying unlike electrical charges). Describe what might happen in some such collisions and sketch corresponding graphs of force versus clock reading. How might you rig a purely mechanical experiment in which the colliding objects stick together instead of springing apart?
- (f) How would a significant amount of friction influence some of the cases you have been observing?
- (g) What do you visualize happens when atoms or molecules collide with each other or with walls of a container?

Note to the student: Question 4.4 illustrates an approach to the idea of conservation of momentum utilizing transformation of frames of reference. This approach provides valuable practice preparatory to study of the theory of relativity. In the latter study you will have occasion to make such transformations repeatedly.

In elastic rectilinear collisions of bodies with equal masses, it is readily observed that (1) when the bodies approach each other with equal and opposite velocities, they rebound with equal and opposite velocities (i.e., they exchange their initial velocities); and (2) when the first body is moving toward the right while the second body is stationary, the first body stops still and the second body moves off with the same velocity as the first (i.e., the bodies again exchange their initial velocities).

Christian Huygens, Newton's great Dutch contemporary in the seventeenth century, raised a penetrating question about these two observations: Are these simply two unconnected and independent phenomena, or is there a deeper order in nature that connects them to some common law or principle? Let us follow his simple and powerful analysis of the problem.

4.4 We start, as Huygens did, with the observation that "equal hard bodies" (meaning perfectly elastic bodies of equal mass) approaching each other with equal and opposite velocities rebound with velocities unchanged in magnitude as illustrated in the following figure. Our frame of reference (set of coordinate axes) is denoted by O. The two bodies, A and B, moving parallel to the x -axis with velocities \vec{v}_{A1} and \vec{v}_{B1} , respectively, undergo a rectilinear collision.

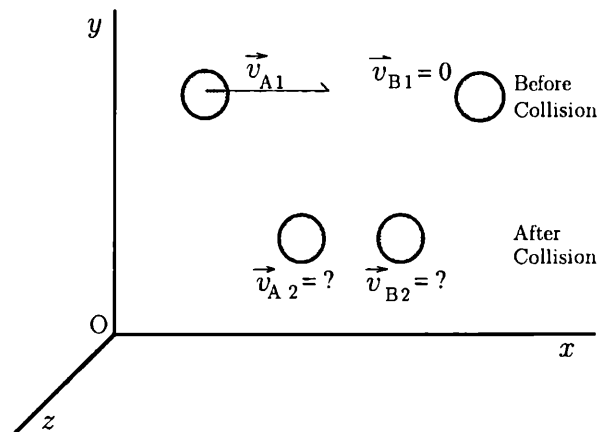


Not only is this starting point consistent with observations but it is also consistent with our deep sense of symmetry in natural phenomena. It is this same sense of symmetry, for example, that leads us to expect identical objects, placed at equal distances from the pivot point of a seesaw, to balance each other. We would be surprised at any other outcome and would be very certain that the two objects were not identical if they failed to balance.

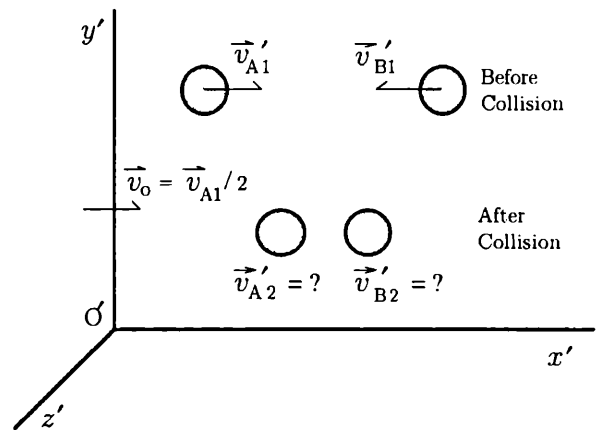
We next ask the question: What would we expect to happen if body B is initially stationary, and an identical body A approaches with velocity \vec{v}_{A1} in the same reference frame O as shown in the next figure.

Huygens proceeded to address this question through the clever device of viewing the second collision from another frame of reference O' , which made the second collision identical in character with the first one. He chose O' to move to the right at uniform velocity $\vec{v}_o = \vec{v}_{A1}/2$ relative to frame O as shown in the succeeding figure.

- (a) Argue in your own words that if we take the velocity \vec{v}_o of frame O' relative to O to be $\vec{v}_{A1}/2$, the two objects, viewed from O' , will appear to be approaching each other at equal and opposite velocities of magnitude $v_{A1}/2$.



- (b) Still from the point of view of O' : The two bodies will now rebound with equal and opposite velocities, also of magnitude $v_{A1}/2$. Enter these after-collision velocities in O' on the succeeding figure, with appropriate symbols.



- (c) Now view these after-collision velocities from the original frame O . Show that from the point of view of O , body A will have zero velocity while body B will be moving to the right with velocity \vec{v}_{A1} .
- (d) Argue in your own words that this analysis shows the two seemingly different types of collision to be intimately related and to be governed by some common underlying order in nature. (This common order turns out to be conservation of momentum, regardless of frame of reference.)

4.5 Carry out an analysis of the perfectly *inelastic* collision of two identical bodies by following a sequence exactly similar to the one used in the preceding problem: Take as initially given the perfectly symmetrical case in which the two bodies, approaching each other with equal and opposite velocities, come to a dead stop on collision. Then consider the problem in which body A moves to the right at velocity \vec{v}_{A1} in frame O

and collides with stationary body B; the two bodies stick together and move toward the right at a final velocity \bar{v}_2 .

- (a) Now view this latter collision from a frame of reference O' in which the two bodies appear to approach each other at equal and opposite velocities and come to dead stop. (What must be the velocity of O' relative to O ?) Sketch the frame and label the velocities as they appear in it.
- (b) Now view the final situation in O' from the original frame O . Show that the velocity of the combination of A and B, stationary in O' , must have the velocity $\bar{v}_{A1}/2$ in O . Note that this constitutes a prediction of what is actually observed to happen.
- (c) What is the point of this analysis? Explain how it reinforces the point made in the preceding question.

4.6 Explain, as though you were making a presentation to your fellow students, the connection between Newton's third law and the law of conservation of linear momentum. Be sure to include the following: An explanation of why it is necessary to define clearly the system under consideration; an explanation of the distinction between an open and a closed system; an explanation of the problem that would arise if the forces the interacting objects exert on each other were not equal and opposite to each other instant by instant throughout the interaction as demanded by Newton's third law. Give an example of some common events in which, at least for a very short time interval, the forces of interaction between well-separated bodies are, in fact, not equal and opposite instant by instant. [Hint: Think of bodies connected to opposite ends of a long, soft spring.]

4.7 Two objects with masses m_A and m_B , respectively, form a closed system and undergo a perfectly elastic rectilinear collision. Object B is initially stationary, and object A has a nonzero initial velocity. Starting with the relevant equations derived in study of such collisions and explaining your reasoning, predict the motion of each body after the collision if

- (a) $m_A < m_B$
- (b) $m_A = m_B$
- (c) $m_A > m_B$
- (d) Now examine and predict what happens in the limit in which m_B is vanishingly small relative to m_A .
- (e) Do the same for the limit in which m_A is vanishingly small relative to m_B .
- (f) Explain the connection between your prediction in part (e) and what happens when a ball bounces elastically from a rigid wall.

4.8 Suppose that in our laboratory frame of reference O_L , a flat, very massive steel wall moves to the right (in the x -direction) with uniform velocity of magnitude v_W . The wall is perpendicular to the x -axis. A small steel ball moves toward the right with a larger velocity v_B , catches up with the wall, and undergoes a perfectly elastic collision, bouncing back toward the left.

- (a) What is the initial velocity of the ball in the frame of reference O_P of the steel plate? With what velocity will the ball rebound in this frame of reference? What will be the final velocity of the ball in frame O_L ? Explain your reasoning.
- (b) Suppose now that the wall moves toward the *left* with uniform velocity v_W as the ball still moves toward the right. What will be the final velocity of the ball in frame O_L after the rebound? Explain your reasoning following a sequence similar to that in part (a).
- (c) In each of the two preceding cases, what happens to the kinetic energy of the ball in frame O_L : Does it increase, decrease, or remain unchanged? Explain your reasoning. If the kinetic energy of the ball changes, where does any decrease go and where does any increase come from?
- (d) A piston, confining a gas in a cylinder, is moving, at uniform velocity, either inward (compressing the gas) or outward (expanding the gas). In the light of your analysis in part (c), describe what must be happening, on the average, to the kinetic energies of molecules of gas that rebound from the piston as the gas is compressed and as it is expanded.
- (e) In the situations visualized in part (d), take the system under consideration to be the gas. Explain where the predicted kinetic energy changes go or come from. (What happens in the way of work being done by the piston on the gas or by the gas on the piston?)

4.9 Suppose we take a toy balloon, blow it up with air, and let it go without tying up its mouth. It is a familiar experience that as the balloon deflates, it flies erratically around the room until it is completely deflated. (In the following, be sure to define the system you will discuss and indicate whether it is open or closed.)

- (a) Without equations or formulas, but using the concepts of impulse and change of momentum, describe, *qualitatively*, the behavior of the balloon while it is deflating.
- (b) Describe the energy changes that take place during the same interval, starting with the situation in the inflated balloon.

4.10 If one wants to jump from a height without getting hurt, one chooses to jump into a stretchable net or pile of hay or pile of soft mattresses rather than hitting the hard ground.

- (a) Explain this choice in terms of the impulse-momentum theorem and the concept of “average force.” How does it come about that the risk of injury is very much less in the softer cases than on the hard ground, while the momentum change is exactly the same in all instances?
- (b) Describe the energy changes that take place in the various circumstances.

4.11 Two objects, one with a large mass m_A and the other with smaller mass m_B , are released from rest relative to an observer in free space. The objects then accelerate toward each other under the influence of their mutual gravitational attraction (no

other forces are acting). Consider the instant just before the objects collide, when they have the final velocities of magnitude v_A and v_B , respectively. (Explain your reasoning in answering each of the following questions.)

- (a) How does the net impulse delivered to A compare with the net impulse delivered to B?
- (b) How does the momentum change of A compare with the momentum change of B?
- (c) How does the final (just before collision) velocity of A compare with the final velocity of B? (Derive an expression for v_A in terms of v_B and the ratio of the two masses.) How do the displacements of the two bodies compare with each other for various ratios of the masses?
- (d) How does the final (just before collision) kinetic energy of A compare with the final kinetic energy of B? Are the two energies equal or is one greater than the other? (Derive an expression for the ratio of the two kinetic energies.)
- (e) Was the work done on A equal to the work done on B? How do you explain the fact that the work done on each body is different while the impulse delivered to each body has the same magnitude?
- (f) Where did the total final kinetic energy of the system of the two bodies come from?

In his *Astronomia Nova* of 1609, seventy-eight years before Newton's *Principia*, Johannes Kepler made what has become a famous and often quoted remark:

If two stones were placed in any part of the world, near each other yet beyond the sphere of influence of a third related body, the two stones, like two magnetic bodies, would come together at some intermediate place, each approaching the other through a distance in proportion to the mass of the other. [In other words, the displacements of the two stones would be inversely proportional to their masses.]

- (g) How does Kepler's prediction compare with your result in part (c)?

Kepler's word for "mass" was then the Latin word "moles"—a term denoting bulk of matter in some vague sense rather than our modern, operationally defined concept. Clearly defined conceptions of inertial mass and gravitational mass were then still far in the future (even beyond Newton's *Principia*). Yet Kepler must be given credit for a profound, partial insight.

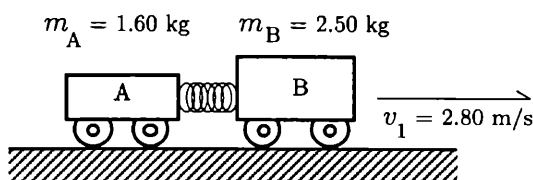
- (h) Comment on Kepler's statement in the light of the analysis you have carried out in the first part of this problem: In modern terms, what are its dynamical implications? Is Kepler's prediction consistent with our present knowledge and concepts? Why or why not? What name do we give to the "intermediate place" to which Kepler refers? What connection, if any, do you see between this situation and the one in which two carts on a level table (or two gliders on an air track) have a compressed spring between them and fly apart when the spring is released?

4.12 A glider of mass m moves in the positive direction on an air track (negligible friction) with velocity of magnitude v_0 . At the end of the track, it makes a perfectly elastic collision with a spring bumper and rebounds with the same magnitude of velocity.

- Explaining your reasoning, draw an appropriate momentum vector diagram, and write an algebraic expression for the change of momentum of the glider.
- What net impulse must have been delivered to the glider by the spring bumper? (Do not lose sight of the fact that impulse and momentum are both vector quantities.) Explain your reasoning.
- What was the kinetic energy *change* of the glider? Explain your reasoning.
- What was the net work done by the spring on the glider? Explain your reasoning.
- Describe the sequence of energy changes that takes place in the glider-spring system between initial contact of the glider with the spring and the final parting of contact.
- Explain how the net work done by the spring on the glider can be zero while the net impulse delivered by the spring to the glider is not zero.
- Suppose now that the collision is partly inelastic and the glider rebounds with a speed of $0.9v_0$. Answer questions (a)-(e) for this case.

4.13 A spring is compressed between two carts and is temporarily clamped so that carts and spring move as a single unit. Initially the carts move toward the right on a level surface with a velocity of 2.80 m/s. The masses of the carts are 1.60 kg and 2.50 kg, respectively, as shown. The mass of the spring is very much smaller than that of either cart and can be neglected. At a certain instant the clamp is suddenly released, and the carts separate, with cart B moving to the right at 3.20 m/s. (In each part of the following analysis, be sure to define the system you are dealing with and to indicate what conservation law or laws you are applying.)

- Indicate your choice of positive direction, and, explaining your reasoning in neat and intelligible sequence, calculate the velocity of cart A immediately after decompression of the spring.



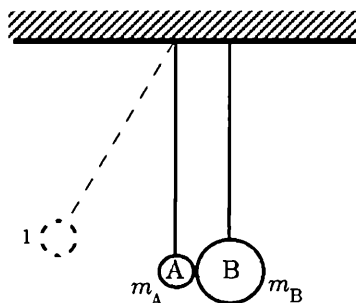
- As a result of this interaction, do you expect the final total kinetic energy of the system (the two carts) to be equal to, greater than, or less than the total initial kinetic energy? Explain your reasoning carefully in terms of the different kinds of energy available in the system. Then check your answer by making the relevant numerical calculations.
- Is this event to be described as involving an elastic or an inelastic interaction between the two carts? Explain your reasoning.

4.14 Suppose you throw a stone having a mass of 0.50 kg vertically upward. Let us assume that your hand exerts an average force of 110 N over an arm displacement (upward) of 0.60 m. (Through the following sequence of questions, we shall explore, in terms of the energy and momentum concepts, what happens to the stone, and we shall ascertain whether the numerical values given above are physically reasonable. Follow the sequence carefully for the exercise it provides in using and interpreting the energy and momentum concepts, setting up numerical expressions and indicating your line of reasoning. Do *not* resort to calculating accelerations or using the *kinematic* relations except to check your results for internal consistency.)

- (a) Draw and label the following force diagrams: For the stone during the act of throwing, for the stone after it has left contact with your hand, for your own body, and for the ground in the vicinity of your feet. Describe each force in words and identify the third law pairs.
- (b) Calculate the work done on the stone *by your hand* in the act of throwing.
- (c) Calculate the *net* work done on the stone during the act of throwing. Explain why this number differs from the one obtained in part (b).
- (d) Calculate the kinetic energy change imparted to the stone in the act of throwing, i.e., the kinetic energy of the stone at the instant it leaves your hand.
- (e) Calculate how high the stone will rise (making use of the kinetic and potential energy concepts).
- (f) Using the result obtained in part (d), calculate the velocity of the stone at the instant it parts contact with your hand.
- (g) Calculate the change of momentum that was imparted to the stone in the act of throwing.
- (h) What magnitude of *net* impulse, in what direction, must have been imparted to the stone in the act of throwing? What total magnitude of impulse was imparted to the stone by your hand? Why is the magnitude of this impulse different from that of the *net* impulse?
- (i) During the act of throwing, were any net impulse and change of momentum imparted to the system consisting of your body and the earth? If not, why not? If so, what was the magnitude and direction?
- (j) Are the numbers given in the problem and the resulting values that you have calculated physically reasonable? Justify your answer on the basis of your own experience in throwing objects upward.
- (k) Where does the kinetic energy imparted to the stone come from?
- (l) How much work is done by the normal force exerted by the ground on your body?
- (m) What is the *rate of change* of its momentum while the stone is on the way up? At the instant it is at the top of its flight? On the way down? Explain your reasoning.

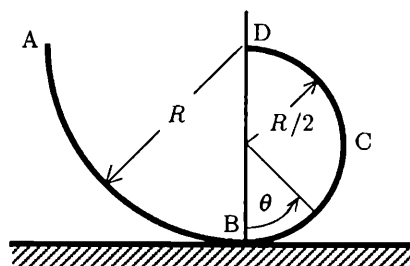
4.15 Suppose two pendulum bobs with known masses m_A and m_B , respectively, are suspended from a rigid support. As implied in the diagram, $m_A < m_B$. Bob A is displaced to position 1 and is let go from rest at that position. The collision between A and B is taken to be perfectly elastic.

- Describe the collision at the bottom of the swing qualitatively: Which way does B move after the impact? Which way does A move? Explain your reasoning.
- Describe how you would proceed to calculate the height to which B rises after the collision: Is there any other information that would have to be given or that you would have to assume? What relations would you use and how would you use them? Is it possible to solve the problem in one step with one single equation or is more than one step necessary? Explain your reasoning in detail in your own words as though you were leading a fellow student through the reasoning and the algebraic steps. Be sure to make clear how the concept of “perfectly elastic collision” is being utilized in the reasoning.
- Is it possible to calculate how high A rises after the impact? If so, how?
- What would you have to know to calculate the initial work put into the system to get it started?



4.16 A small cart starts from rest at point A and rolls down a circular track of radius R . At point B (the lowest point on the track), it enters another circular track, which has a radius $R/2$. Positions of the cart along this latter track can be described in terms of the angle θ measured from the vertical line through B. In the following analysis, treat the cart as a particle moving along arcs of the radii indicated and the motion as frictionless.

- Think carefully about how the normal force N exerted by the track on the cart beyond point B must be varying as the cart ascends the track and argue qualitatively (without mathematical analysis) that the cart cannot drop off the track before it reaches point C and that it must drop off the track before it reaches point D.



- Now proceed to confirm the qualitative argument in part (a) with a mathematical analysis: Using a conservation of energy argument, show that the square of the velocity v_B of the cart at point B is given by

$$v_B^2 = 2gR$$

(Be sure to define the system to which you are applying the energy argument.)

- (c) Using a similar conservation of energy argument, show that beyond point B, the square of the velocity of the cart depends on θ in the following way:

$$v_B^2 = gR(1 + \cos \theta)$$

- (d) Draw a force diagram for the cart at an arbitrary θ and apply Newton's second law to the motion to solve for the normal force N exerted by the track on the cart as a function of θ . Interpret the equation for N , showing that the result confirms the simple argument made in part (a), and find the numerical value of the angle θ at which the cart will drop off the track. [Hint: What value of N will signal the parting?] (Answer: $\theta = 120^\circ$) Point out the terms that would be affected by the presence of friction and show that the effect would be to decrease the angle at which the cart drops off, as one would expect intuitively.
- (e) How does it come about that even in the frictionless situation, the cart cannot get back to its original height R above point B? Is this a violation of conservation of energy? Why or why not?
- (f) Identify some kinematic and dynamic variables that change abruptly at point B. Identify some variables that do not change abruptly at point B.

4.17 Lord Kelvin describes encountering Joule and his bride honeymooning in Switzerland. Joule was stopping by roadside waterfalls and measuring temperatures of the water in the stream entering the falls and of the water in the still pool at the bottom of the falls.

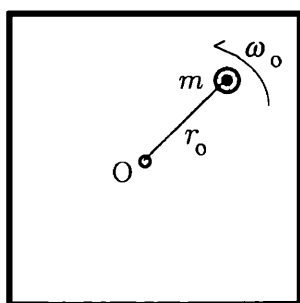
- (a) Describe in good English the energy transformations that occur between the initial and final conditions in which Joule was interested. From what you know about Joule's work and interests, why do think he was bothering with these observations?
- (b) Estimate the temperature difference that might arise ideally under these circumstances in the case of a waterfall 50 m high. Explain your reasoning carefully.
- (c) Is your estimate of temperature likely to be an upper limit, a lower limit, or something in between? Explain. List the effects that would make the actual temperature difference deviate from the ideal result. What would be the direction of the deviation? Explain your reasoning.
- (d) How sensitive a thermometer would you need if you wished to make similar observations? Is such a thermometer readily available?

4.18 A satellite moves around the earth in a sufficiently large circular orbit that the frictional resistance to the motion is essentially zero. Explain how we know that under these circumstances, the motion will persist without our putting any additional energy into the system.

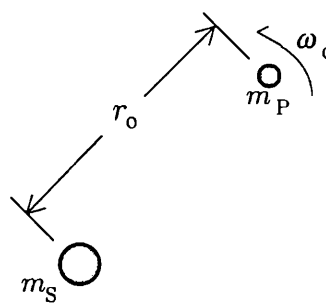
4.19 Consider a planet orbiting the sun in an elliptical orbit. Sketch such an ellipse, label some key positions, and then describe *qualitatively* how the kinetic energy of the planet and the potential energy of the planet-sun system vary as the planet executes its orbit.

4.20 Let us analyze and compare the energy changes that take place in systems 1 and 2: The first consists of a frictionless puck of mass m on an air table. The puck is attached to a string, which can be pulled down through a smooth hole in the air table at point O. The puck is set into circular motion at angular velocity ω_o about point O at an initial radius r_o .

System 2 consists of a planet with mass m_P in circular orbit at angular velocity ω_o and at radius r_o around a very massive sun with mass m_S . (With m_S very much larger than m_P , any motion of the sun will be completely negligible, and we need concern ourselves only with the motion of the planet.)



System 1



System 2

Suppose the string in system 1 is pulled slowly down through the hole so that the frictionless puck is drawn in from the initial radius r_o to a final smaller radius r and corresponding angular velocity ω without acquiring appreciable kinetic energy in *radial* motion. This is like raising an object slowly without imparting appreciable kinetic energy. In raising the object, we calculate the work we do on the object-earth system by taking the force we exert to be essentially equal in magnitude to the weight of the object. In the case of the frictionless puck, the force we exert in pulling the string is essentially equal to the centripetal force instantaneously exerted on the object. (The situation is analogous to the one in which the spinning skater draws in his or her arms and experiences increasing angular velocity while angular momentum is being conserved.)

- (a) Treating the puck as a point particle and carrying out the necessary integration, obtain an algebraic expression for the work W done on the puck in displacing it from the initial radius r_o to a final radius r subject to the auxiliary condition that the angular momentum $mr^2\omega$ of the puck is conserved. You will find that the final expression can be written in either of two ways:

$$W = \frac{1}{2} m r_o^2 \omega_o^2 \left[\frac{r_o^2}{r^2} - 1 \right] \quad (1)$$

$$W = \frac{1}{2} m r_o^2 \omega_o [\omega - \omega_o] \quad (2)$$

- (b) Given the initial and final conditions for r and ω as defined above and the fact that the instantaneous tangential velocity $v_t = r\omega$, argue, independently of the relations (1) and (2) for W , that the change in kinetic energy ΔKE of the puck

must be given by

$$\Delta KE = \frac{1}{2} m r^2 \omega^2 - \frac{1}{2} m r_o^2 \omega_o^2 \quad (3)$$

- (c) Now show that either one of the expressions (1) or (2) for W is in fact equal to expression (3) for ΔKE . How do you interpret this equality? That is, what happened to the work done in pulling the string and decreasing the radius from r_o to r ? Does this make sense in terms of energy conservation? Why or why not?
- (d) Now suppose that in system 2, the planet is slowly lowered in toward the sun just as the puck was pulled in toward the center of its circle. (This is like slowly lowering an object in the earth's gravitational field.) Show that the work that must be done on the planet-sun system to effect this displacement is given by

$$W = -G m_P m_S \left[\frac{1}{r} - \frac{1}{r_o} \right] \quad (4)$$

Explain the meaning of the minus sign and argue that this expression is, in fact, the decrease in potential energy of the planet-sun system for the specified change in radial position of the planet.

- (e) Treating the planet as a point particle, argue that the change ΔKE in its kinetic energy must be given by an expression identical to (3) for the frictionless puck, simply replacing m by m_P . Then show that this expression can be reduced to the form

$$\Delta KE = \frac{1}{2} G m_P m_S \left[\frac{1}{r} - \frac{1}{r_o} \right] \quad (5)$$

- (f) Note that, in the case of system 2, the work done on the planet-sun system in changing the radial position of the planet is *not* equal to the change in the kinetic energy of the planet, while in system 1 the work done in changing the radial position of the puck *is* equal to the change in the puck's kinetic energy. How do you account for this profound difference between the two systems? Is this a violation of the law of conservation of energy? Why or why not? (In the gravitational case, what must happen to the difference between the decrease in potential energy of the system and the increase in kinetic energy of the planet?) Under what circumstances, in the gravitational case, might the radial position of a planet or satellite be caused to increase or decrease? (In your comparison of systems 1 and 2, note that in system 1, the string exerts a passive force that adjusts itself to the given conditions and that any angular velocity is possible at any radius as long as the string doesn't break. However, in system 2, the law of gravitation requires that a given value of angular velocity is possible at only one particular radius.)
- (g) Compare system 2 with the situation in which a small negatively charged particle with mass m_P and charge $-q_P$ revolves in a circular orbit around a massive positively charged particle with mass m_S and charge $+q_S$. Discuss what effects might lead to an increase or decrease in the radial separation in this system.

- (h) Return to Question 3.37 if you worked on it previously. How do you account for the fact that the satellite, under the influence of a drag force, gains, rather than loses, kinetic energy?

4.21 Suppose that two carts, with a compressed spring between them, are free to roll on a level table. The carts have very different masses and are initially stationary; consider friction to be negligible. The spring is released, and the carts fly apart. How do the kinetic energies of the carts compare at the instant they part contact with the spring? That is, are the kinetic energies equal or is one larger than the other? To answer this question, choose your own symbols and analyze the situation algebraically, explaining your reasoning and interpreting the final result.

4.22 A coil spring obeying Hooke's law has a spring constant denoted by k . Suppose, by applying a force denoted by F , we stretch the spring by a length ΔL from its relaxed position.

- (a) Sketch a graph of applied force versus spring extension from 0 to ΔL .
- (b) Appealing to the definition of "work," explain why areas under this graph, for different extensions ΔL , represent amounts of work done on the spring by the applied force. Explain what happens to the work that is done on the spring.
- (c) Shade on the graph an area that represents the amount of work done in stretching the spring from extension 0 to $\Delta L/4$. Use a different shading to indicate the area that represents the work done in stretching the spring from extension $\Delta L/2$ to $3\Delta L/4$. Explain why the amount of work done in the second case is different from that done in the first, even though the change of length of the spring is the same in both cases.

4.23 A pendulum bob on a string of length L is elevated from its lowest position to the point at which the string makes an angle of 90° with the vertical. Take the potential energy of the bob-earth system to be zero when the bob is at this level. The bob is now released, swinging from the 90° to the 0° position.

- (a) Compare the magnitude of change in potential energy of the system between the 90° position and the 45° position with the change between the 45° position and the 0° position. Is one change larger than the other or are they equal in magnitude? Explain your reasoning.
- (b) How much work is done by the string on the bob during the descent? Explain your reasoning.

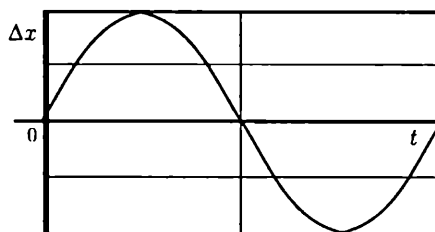
4.24 A ball is thrown from ground level with initial horizontal velocity component v_{0x} and initial vertical velocity component v_{0y} and returns to ground level. Neglecting friction and explaining your reasoning in each instance, write expressions in terms of these two velocities for:

- (a) The largest kinetic energy of the ball during its flight.
- (b) The smallest kinetic energy of the ball during its flight.

- (c) The maximum potential energy of the ball-earth system during the flight.

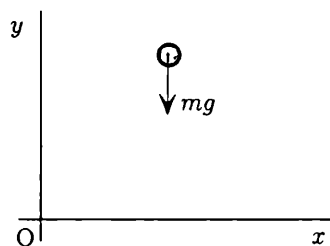
4.25 The following sine function graph shows one complete cycle of the displacement-clock reading history of a simple harmonic oscillation of a block on a spring.

On a similar set of coordinates directly below this graph, sketch the corresponding history of kinetic energy versus time for the block. (This is a purely qualitative question. No numbers are available; just sketch the shape of the required graph.) Give a brief description of the reasoning you used in making the sketch.



Note to the instructor: Questions 4.26 through 4.28 are designed to help students master the meaning and interpretation of algebraic signs that arise in describing the energy changes that take place in several simple cases. Such mastery of algebraic signs can be attained only through practice of this kind, and the practice is rarely available.

4.26 Let us consider an object (say a ball) that moves up or down freely in the vertical direction. Air resistance is assumed to be negligible; only the gravitational force is acting; and we take positive direction upward as shown. We treat the symbol g for the acceleration due to gravity as a magnitude only and not as a vector quantity with directional algebraic signs.



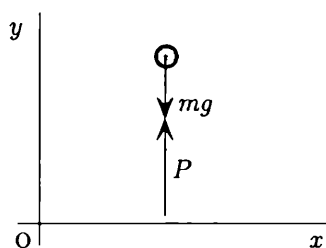
- (a) Show that the work-kinetic energy theorem applied to a vertical displacement Δy in this situation takes the form

$$-mg \Delta y = (1/2)m v_2^2 - (1/2)m v_1^2$$

- (b) Now consider what happens when you let the ball go from rest at some initial high elevation y_1 . What is the value of v_1 ? Which is the only direction in which the ball will move? What will therefore be the algebraic sign of any possible Δy ? What will be the algebraic sign on the left-hand side of the equation? What does the equation say must happen to the kinetic energy of the ball under these circumstances? Does this make sense? Why or why not?
- (c) Suppose, instead of letting the ball go from rest, we give it an initial *upward* velocity v_1 and therefore an initial kinetic energy $(1/2)m v_1^2$. What will be the sign of Δy and the algebraic sign on the left-hand side of the equation? Examine what the equation says must now happen to the kinetic energy as the ball rises. What does the equation say about how large Δy can become? (Argue that the result loses physical meaning after a certain value of Δy .) What happens after the largest Δy is attained?

- (d) Now suppose that the x -axis represents a floor with which the ball can undergo a perfectly elastic collision, reversing its velocity on impact with no loss in magnitude. Show that the equation above says that if the ball is released from rest at some initial height above the floor, it will continue bouncing up and down, always returning to the height at which it was released.
- (e) Suppose that the collision with the floor is not perfectly elastic and that the magnitude of the velocity decreases somewhat on each bounce. Show what the equation says will happen under these circumstances.

4.27 Let us consider the case in which an object (say a ball) moves up or down in the vertical direction under the combined influence of its weight mg and a constant upward force of magnitude P as shown. P might be equal to mg , or larger or smaller. The imposed vertical displacement is denoted by Δy . Air resistance is assumed negligible, and we take positive direction vertically upward.



We treat the symbol g for the acceleration due to gravity as a magnitude only, not as a vector quantity with directional algebraic signs. In the following analysis, we shall take P , mg , and Δy as specified quantities and ask what happens to the kinetic energy of the ball under various circumstances.

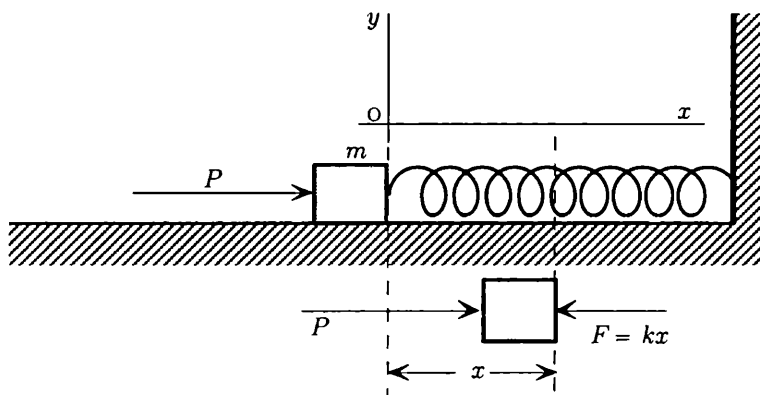
- (a) Show that the work-kinetic energy theorem applied to this situation takes the form

$$(P - mg)\Delta y = (1/2)mv_2^2 - (1/2)mv_1^2$$

- (b) Suppose first that P is *larger* than mg and that Δy is positive. Which way has the ball been moving during the displacement being considered? What is the algebraic sign of the left-hand side of the equation? What must be the final kinetic energy of the ball (in terms of the known quantities) if it started from rest at position y_1 ? What must be its final kinetic energy if it had velocity v_1 at position y_1 ? Show that the final algebraic results correspond to what you would expect to happen physically in these circumstances.
- (c) Suppose that P is *smaller* than mg and that Δy is negative. Which way has the ball been moving during the displacement being considered? What is the algebraic sign of the left-hand side of the equation? What must be the final kinetic energy of the ball (in terms of the known quantities) if it started from rest at position y_1 ? What must be its final kinetic energy if it had velocity v_1 at position y_1 ? Show that the final algebraic results correspond to what you would expect to happen physically in these circumstances.
- (d) Suppose that P is *smaller* than mg but that Δy is positive. Could the ball have started from rest at position y_1 ? Why or why not? Which way has the ball been moving during the displacement being considered? What is the algebraic sign of the left-hand side of the equation? What must be the final kinetic energy of the ball (in terms of the known quantities) if it had an upward velocity v_1 at position y_1 ? Show that your final algebraic result corresponds to what you would expect to happen physically in these circumstances.

- (e) Suppose that P is larger than mg but that Δy is negative. Examine and interpret this situation in the manner that has just been outlined in part (d).
- (f) Discuss the situation in which P is just infinitesimally greater or smaller than mg so that the ball is displaced up or down but with negligible acceleration and essentially zero velocity.

4.28 A frictionless puck of mass m rests on a level air table. The puck is connected to one end of a spring that has a spring constant k and a mass that is negligible compared to the mass of the puck. The other end of the spring is fastened to the wall as shown in the diagram. The origin of x coordinates is located at the relaxed position of the left-most end of the spring. In the following analysis and visualizations, let us confine the puck to displacements that do not damage or exceed the capacity of the spring. The free body force diagram of the puck has omitted the balanced vertical forces so as to avoid cluttering the figure.



A horizontal force P , which may vary with position, is exerted on the puck, displacing it to the right. The spring obeys Hooke's law, and the opposing force, kx , exerted by the spring on the puck increases linearly with displacement from the relaxed position.

- (a) Show that if the force P is applied when the puck is initially at rest at the origin, the work-kinetic energy theorem requires that

$$\int_0^x P(x) dx - \frac{1}{2} k x^2 = \frac{1}{2} m v^2 \quad (1)$$

where x denotes the final position and v the velocity at that position.

- (b) Interpret what Eq. (1) is telling us. For example, if the left-hand side of the equation is positive, what must have been happening in the way of motion of the puck? What happens to the net work done on the puck? What does the puck do if the force suddenly drops to zero at position x ?
- (c) Show that for the more general case in which the force P is applied between an initial position x_1 and a final position x_2 , the work-kinetic energy theorem requires that

$$\int_{x_1}^{x_2} P(x) dx - \left[\frac{1}{2} k x_2^2 - \frac{1}{2} k x_1^2 \right] = \frac{1}{2} m v_2^2 - \frac{1}{2} m v_1^2 \quad (2)$$

- (d) Interpret Eq. 2. What is happening if the left-hand side of the equation is positive? If the left-hand side is negative?
- (e) Show that if the puck is released from rest ($v_1 = 0$) at some initial position x_1 not at the origin and with the force $P = 0$, Eq. 2 requires that from then on

$$\frac{1}{2}kx_1^2 = \frac{1}{2}kx_2^2 + \frac{1}{2}mv_2^2 \quad (3)$$

- (f) Interpret Eq. 3 and contrast the variation of the kinetic energy of the puck on the spring (when released from rest) with the variation of the kinetic energy of the ball released from rest in free fall. Show that very similar conservation relationships obtain, but that there are important differences in the character of the motion that can take place. If we drop a body vertically, from rest, through a displacement Δy , it acquires a kinetic energy $(1/2)mv^2$ equal to the magnitude of $mg\Delta y$ and will continue dropping and acquiring still more $(1/2)mv^2$ if the fall is not interrupted. If in the case of the spring, however, we release the puck from rest at a position x_1 , it will be accelerated toward the left, and, on returning to $x = 0$, it will have acquired an amount of kinetic energy equal to $(1/2)kx^2$. It will then continue moving to the left, but the acceleration will now be directed to the right because of the stretching of the spring. In the case of free fall, there was no such reversal of the acceleration. Show that the equations tell us that the puck will continue moving to the position $x = x_1$, at which point the direction of motion will be reversed. In the absence of friction, this oscillation would continue indefinitely (as in the case of the bouncing ball) with the maximum values of $(1/2)kx^2$ and $(1/2)mv^2$ being continually interchanged. The motion is "symmetrical" around the position $x = 0$. What mathematical characteristic of Eq. 3 accounts for this symmetry? Solve Eq. 3 for v_2 and interpret the result very carefully in words.

4.29 Look around the room or place in which you happen to be located. Identify and describe qualitatively at least two or three processes of transformation of energy that are going on right now in your vicinity. Do not ignore the fact that your own body may be involved in some of the processes. Your description should be detailed in the sense of defining and describing relevant systems, changes in state, interactions, forms of energy involved, and forms of energy transfer that are taking place. Note that there is no possibility whatsoever that no energy transformations are taking place. We ourselves and everything around us are immersed in a ceaseless flux of energy transformation.

4.30 When you are at the top of a staircase, a certain amount of gravitational potential energy is stored in the system consisting of your body and the earth. Describe what happens to this potential energy as you descend the staircase. Be sure to indicate clearly what other objects or systems become involved in the associated interactions and energy transfers and what forms of energy are involved.

4.31 The energy transformations taking place in chemical reactions (e.g., the amount of heat received from or transferred to surrounding systems) depend quantitatively on the initial and final states of the objects or materials constituting the reacting system. The initial state is defined by many properties or factors such as temperature,

pressure, composition, internal stresses and their attendant deformations, and the effects of imposed electric and magnetic fields.

In the light of the foregoing statement, speculate on what happens to the stored potential energy when a compressed metallic spring is dissolved in acid and seems to disappear completely as far as its initial state, appearance, and configuration are concerned. Has the stored potential energy been “destroyed,” thus violating the conservation law? In what way is it likely that energy is still being conserved in these circumstances? (Compare this situation with what might be happening when the same *uncompressed* spring is dissolved in acid.)

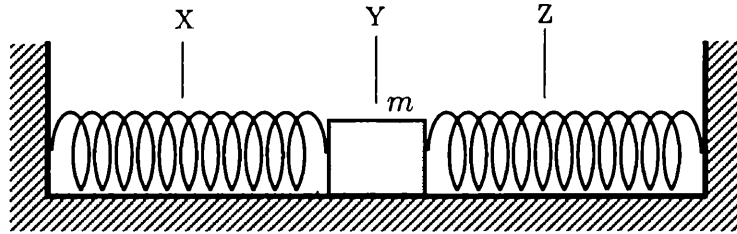
Note to the student: In questions 4.32 and 4.33, circle the letters marking all those statements that are correct. Any number of statements may be correct, and therefore each one must be examined on its merits. Do not simply drop the question after you have found one correct statement.

4.32 A block of mass m is pulled along the floor by a force T inclined at an angle θ as shown in the diagram. The coefficient of friction between the block and the floor is denoted by μ . The magnitude of the force is such that the block moves with uniform velocity.



- (a) The magnitude of the normal force exerted on the block by the floor is given by:
- | | |
|--------------------------|--------------------------|
| (A) mg | (D) $mg - T \sin \theta$ |
| (B) $mg - T \cos \theta$ | (E) $T \sin \theta$ |
| (C) $mg + T \cos \theta$ | (F) None of the above |
- (b) The magnitude of the frictional force exerted on the block by the floor is given by:
- | | |
|---------------------|-------------------------------|
| (A) $T \cos \theta$ | (D) $\mu(mg - T \cos \theta)$ |
| (B) $T \sin \theta$ | (E) Zero |
| (C) mg | (F) None of the above |
- (c) During a horizontal displacement Δx of the block:
- (A) The work done by the force T is given by $T\Delta x$.
 - (B) The work done by the force T is given by $(T \sin \theta)\Delta x$.
 - (C) Some of the work done by the force T is stored as potential energy in the system while the rest is converted into thermal internal energy.
 - (D) The work done by the force T is equal to the kinetic energy change of the block.
 - (E) The change in potential energy of the block-earth system is $mg\Delta x$.
 - (F) None of the above.

4.33 A frictionless puck of mass m , mounted between identical springs as shown, can slide back and forth on the level frictionless surface. The springs have negligible mass relative to the mass of the puck.



The puck is displaced by hand from its equilibrium position at Y to position X, at which point it is released from rest. It then oscillates back and forth between positions X and Z. Circle the letters marking the correct statements about the oscillatory motion.

- (a) The puck has its largest value of kinetic energy at position Z.
- (b) The puck has its largest value of kinetic energy at position Y.
- (c) The system has its largest potential energy when the puck is at position Y.
- (d) The system has its largest potential energy when the puck is at position Z.
- (e) The potential energy of the system when the puck is at position Z is equal to the work that was done in displacing the puck from Y to X.
- (f) The direction of momentum change of the puck is toward the right throughout any interval in which the puck is located between Y and X regardless of which way it is moving.
- (g) The rate of change of momentum at position Y is zero.
- (h) The rate of change of momentum has its largest magnitude at positions Y and Z.
- (i) At any instantaneous position of the puck, the sum of the instantaneous values of kinetic energy of the puck and potential energy of the system is equal to the work that was initially done in displacing the puck from Y to X.
- (j) None of the above.

Note to the student: Questions 4.34 through 4.36 describe a physical situation and then make a statement (or statements) about it. Accept the description of the physical situation (the numbered item) as given and correct. Examine the statements, including numerical values, to determine whether they are correct. If they are correct, say so explicitly. If they are incorrect, alter them to eliminate errors.

4.34 A frictionless puck on a string on a level air table moves in a horizontal circle at uniform angular velocity.

- (a) The string exerts a force on the puck, and the work done by this force is equal to the rotational kinetic energy of the bob.

- (b) The linear momentum of the puck does not vary as the puck continues in uniform circular motion.
- (c) The force exerted by the string keeps imparting an impulse of constant magnitude to the puck.
- (d) The vertical force exerted by the table on the puck delivers zero impulse to the puck.
- (e) The potential energy of the system consisting of the puck and the string can be taken to be zero.
- (f) The system consisting of the puck, the string, and the table can be regarded as a closed system whether or not friction is present.

4.35 A ball with a mass of 250 g is thrown vertically upward. It rises to a position 10 m above the point at which it left the thrower's hand. Neglecting frictional effects:

- (a) The velocity with which the ball left the thrower's hand must have been about 14 m/s.
- (b) The upward impulse delivered to the ball in the act of throwing must have been about 1.4 N s.
- (c) The downward impulse delivered to the earth in the act of throwing must have been zero.
- (d) The kinetic energy imparted to the ball must have been about 2.5 J.
- (e) The work done by the thrower must have been about 2.5 J.

4.36 Two identical gliders move toward each other with equal speeds on a level air track.

- (a) The total momentum of the system consisting of the two gliders is zero.
- (b) The total kinetic energy of the system consisting of the two gliders is greater than zero.
- (c) The kinetic energy of the system consisting of the two gliders will be reversed after the gliders have undergone perfectly elastic collision.
- (d) With no external horizontal forces acting on the system consisting of the two gliders, the net impulse delivered to each one of the gliders must be zero over the interval of collision.

Chapter 5

Electricity

5.1 What is meant by the term “electrostatic interaction”? Describe the circumstances to which this name is applied. We also speak of “magnetic interaction” and “gravitational interaction.” How do we distinguish one of these interactions from another? What are the similarities and what are the differences? (In answering this question, it is necessary to describe specific instances of what does and does not happen in these various phenomena. The larger your list the better the answer.)

5.2 (a) What is meant by the term “electrical charge”? Describe observed effects that lead us to invent this property for objects that have been treated in an appropriate way, even though we have no idea of what charge “is.” (b) Describe observations that lead us to infer that charge, whatever it might be or however it might be carried, is movable or transportable from one object to another. (c) Describe observed effects that are interpretable in terms of the concept of altering the “quantity of electrical charge.” (d) Describe observed effects that lead to the inference that the *strength* of electrical interaction (i.e., the force exerted by charged objects on each other) depends on at least two variables: Separation between the objects and quantity of charge carried by each.

5.3 What is meant by the term “like” when we talk about electrical charges? How does this term originate? Answer the question by describing experiments and what is actually observed. What is the origin of the statement “like charges repel”? What is “alike” about like charges: Do they look, feel, sound, smell similar? Is the introduction of the word “like” a matter of arbitrary definition? Could we have equally well started with the definition “unlike charges repel” and based our system on that terminology, or do observations *dictate* the familiar terminology?

5.4 On what basis do we find that we can get away with only the two terms “like” and “unlike”? In other words, what is the basis for believing that no more than two varieties of electrical charge exist in nature even though we observe electrical interaction among a vast variety of different materials and in different circumstances? To make your argument convincing, you will need to describe observations or experiences (visualized in the abstract) that would force you to recognize and accept a third variety of electrical charge if it happened to turn up on a new material. The fact that no such interaction has ever been observed leads to the deeply held belief that there are no more than two varieties of charge. (Note that understanding the significance of what does *not* happen is sometimes just as important as knowing what

does happen.) Where do the terms “positive” and “negative” come from? Are they necessary? Would other names do just as well?

5.5 What observed effects lead to the inference that both varieties of electrical charge are initially present in all objects in balanced amounts? What observed effects lead to invention of the model to which we give the name “polarization,” visualizing the shifting or displacement of charges relative to one another in one object when another charged object is brought near.

5.6 Describe observations that lead to the invention of the concepts “conductor” and “nonconductor” or “insulator.”

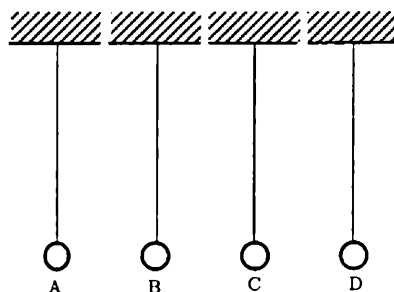
5.7 A hard rubber or plastic rod is rubbed with fur and is touched to an electroscope. The electroscope exhibits the presence of a net charge with the flexible leaf standing away from the fixed center post. If one brings one’s hand or any large uncharged object near the electroscope platform, the leaf is observed to drop somewhat.

Interpret this decrease in deflection by describing what we visualize as happening in the way of charge displacements in both the electroscope and the uncharged body. Do this by drawing diagrams showing the altered charge distributions and explaining your reasoning. How do you account for the fact that the shifting of charge does not simply go on indefinitely but that equilibrium configurations are attained?

5.8 A toy balloon, after being rubbed with a cloth or sheet of plastic, is brought up to the wall of the room and sticks to the wall without being held or otherwise supported. Explain in detail what is happening: Sketch the electrical effects that arise (not just in the balloon but also in the wall). Draw well-separated force diagrams for the balloon and for the region of the wall it touches while it is sticking to the wall. Describe each force in words and identify the third law pairs. What holds the balloon vertically? Would the behavior of the balloon be different if it were hard and round instead of soft and deformable? If so, how would it behave?

5.9 Suppose we have four conductor-coated pith balls A, B, C, and D suspended on nonconducting (e.g., nylon) threads. The charge states of all four balls are initially unknown. We now take a rubber rod, rub it with fur so that it becomes negatively charged, and bring it in contact with sphere A. Then we bring the balls near each other (without contact) two by two. The interactions are as follows: (1) B, C, and D are all attracted to A; (2) B and C have no discernible effect on each other; (3) B and C are both attracted to D.

- (a) What are the charge states of A, B, C, and D? Explain your reasoning.
- (b) In the story above and its conclusions, what were the observations and what were the inferences?



Suppose, by handling sphere C with the nylon thread, we now bring it close to (but not touching) sphere A. While the two spheres are close together, we touch C briefly with one finger. Then we remove C from the vicinity of A.

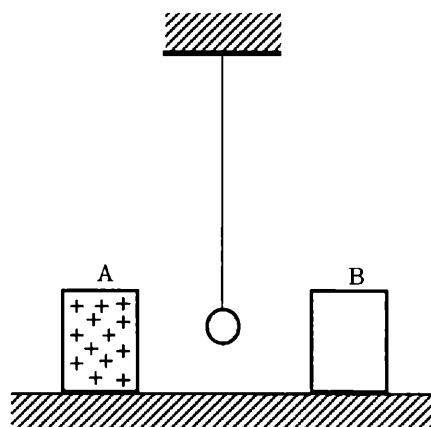
- (c) What is now the charge state of C? Explain your reasoning, describing, with words and pictures, what happens to C step by step through the entire sequence. What is the name for the phenomenon you have just described?
- (d) How will C now interact with A, B, and D? Explain your reasoning.

5.10 Two identical conducting spheres A and B carry equal amounts q of like charge and, separated by a distance that is large compared to their diameters, repel each other with a force of magnitude F . A third identical conducting sphere D is mounted on an insulating handle and can be moved around at will by the experimenter. The experimenter first brings the initially uncharged sphere D in contact with sphere A, and then, without discharging D, brings it in contact with sphere B. Sphere D is then removed from the vicinity of A and B.

- (a) How will the magnitude of the force with which A and B now repel each other compare with the initial value F , the distance between A and B remaining unaltered? Give a numerical value for the ratio and explain your reasoning.
- (b) What will now be the amount of charge on sphere D? Give the answer in terms of a numerical factor times q and explain your reasoning. (Be sure to verify that charge has been conserved overall.)

5.11 Two metal cans A and B are placed near each other on a table as shown. A pith ball on a nylon string is suspended so that it hangs between the two cans. Can A is now charged positively by contact with a positively charged rod. The pith ball is attracted to the charged can, makes contact, swings over to can B, makes contact, flies off, and then continues to oscillate back and forth between the two cans.

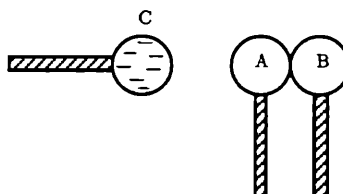
- (a) Sketch a series of diagrams in which you show: (1) How the uncharged pith ball initially becomes attracted to the charged can. (2) What happens when it makes contact with the charged can. (3) Why it swings over to the other can and what happens when it makes contact there. (4) Why it continues to swing back and forth. Accompany each of your diagrams with an explanation of what is happening. (You should be able to carry out this description in any one of the following ways: (1) Conventional displacement of positive charge with negative charge fixed. (2) Displacement of negative charge with positive charge fixed. (3) Both varieties of charge displaced simultaneously.)



- (b) How long will the back-and-forth oscillation continue? What is the criterion for its stopping?

5.12 Metal spheres A and B, standing on insulating supports, are in contact with each other. Another sphere C, highly charged negatively through contact with a Van de Graaf generator or a Wimshurst machine, is brought near A and B as shown. While C is nearby, B is moved off to the right so that A and B are now separated. C is then removed from the vicinity.

- (a) A and B are now charged. Describe, with the help of appropriate diagrams, how they became charged, and identify the kind of charge present on each. What is the name of the process by which B became charged?

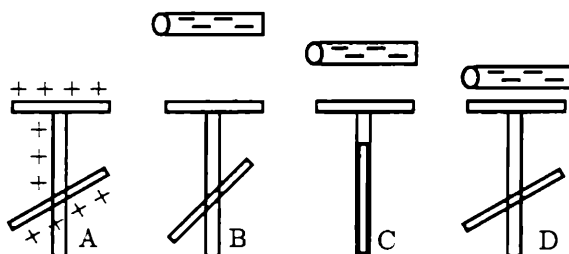


When B is now brought back near A (without contact), a spark is likely to jump between the spheres with an obvious appearance of light, heat, and sound.

- (b) How did it become possible for energy to be released in this manner? Describe the energy transformations that took place in the entire sequence outlined.

5.13 In the sequence illustrated in the following diagram, A shows an electroscope carrying a net positive charge and exhibiting a deflection of its needle accordingly. (Remember that in the convention we use to indicate charging, we show only the *excess* variety of charge in various regions, not the underlying internal sea of balanced, distributed positives and negatives.)

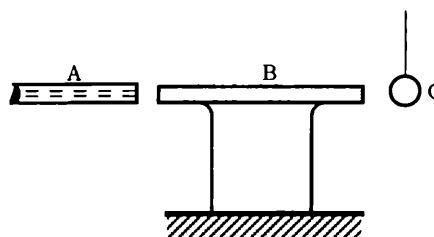
A negatively charged rod is now brought toward the electroscope in three successive steps, somewhat closer in each step, as illustrated. In B the electroscope shows somewhat less deflection than in A. In C the deflection is just zero. In D the deflection is again large.



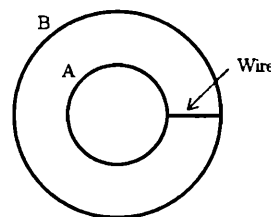
In each of the diagrams B, C, and D, sketch an arrangement of excess positive or negative charges in various regions to show what happens on the electroscope as the charged rod is brought closer and the deflection of the needle changes.

5.14 An uncharged metal rod B rests on a glass beaker as shown. A conductor-covered pith ball C hangs on a nylon thread near the right-hand end of the rod. When a strongly negatively charged plastic rod A is brought near the left-hand end of B (without making contact with B), the pith ball is attracted to the rod, makes contact with the rod, and then flies off. The plastic rod is then removed, still without having touched B.

With an appropriate sequence of sketches of your own choice and with accompanying verbal description, show what happens in this system, stage by stage, with respect to charging and charge distributions, and predict the final charge condition of both rod B and pith ball C.



5.15 Consider the situation in which a hollow conducting sphere A is located *inside* another hollow conducting sphere B. Sphere B has twice the diameter of A. The two spheres are connected by a metal wire as shown. A quantity of charge Q is transferred to sphere B by contact from an electrostatic generator. What fraction of the charge Q is transferred to sphere A as equilibrium is attained in the system? Explain your reasoning.



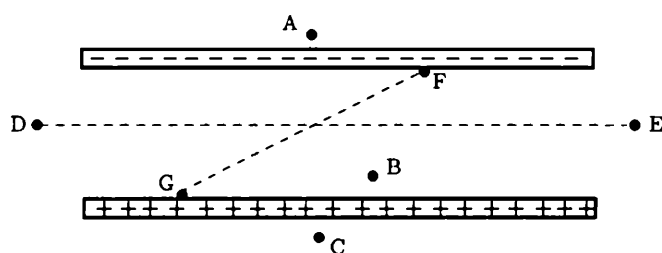
5.16 Given the fact that two charged metallic conductors are in equilibrium with respect to transfer of charge from one to the other when their potentials relative to infinity are equal. Given also the fact that the potential of an isolated, charged metal sphere is inversely proportional to its radius R .

Now consider the case in which an uncharged sphere A with radius R_A is allowed to come to equilibrium with a charged sphere B with radius R_B , carrying a quantity of charge Q . Since bringing the spheres into direct contact, surface to surface, results in drastic alteration of the charge distribution and may do complicated things to the potential, let us perform the “thought experiment” of achieving equilibrium by keeping the two spheres fairly widely separated and transporting charge from one to the other by moving a very small conducting sphere back and forth between them until no more charge is transferred. Under these circumstances, the two principal spheres maintain reasonably well-defined potentials, and equilibrium is attained when the two potentials are equal. (Still another thought experiment approximating the same situation might be the connecting of the two spheres by a long, thin wire.)

- Explaining your reasoning, obtain expressions for the final quantities of charge q_A and q_B carried by each of the spheres at equilibrium. Give the expressions in terms of the radii and Q , the initial total charge on sphere B.
- Interpret the algebraic expressions you have obtained. How is the charge distributed when the two spheres are identical? What happens to the charge received by A as R_A is made very much smaller than R ? As R_A is made very much larger than R_B ?
- Consider the following assertion: “When R_A is very small relative to R_B , the charge received by sphere A is actually a close measure of the initial potential of sphere B; when R_A is very much larger than R_B , the charge received by A becomes a measure of the quantity Q initially carried by B.” Is this assertion correct or incorrect? Explain your reasoning.
- To a rough approximation, an ordinary electroscope might be treated in terms of the ideas developed above. In the light of the conclusions reached in part

- (c), what would you say the deflection of an electroscope leaf measures (approximately) when the electroscope makes contact with an object very small relative to itself? When the electroscope makes contact with an object very large relative to itself?
- (e) Suppose the initially charged hollow sphere B is much larger than sphere A and has a hole in its surface so that objects on handles can be inserted inside the spherical shell. Uncharged sphere A, carried on an insulating handle, is inserted into sphere B and brought in contact with the inside surface of B. The initial charge of B is still Q , and the radii are R_A and R_B , respectively. What is the final equilibrium distribution of charge between A and B under these circumstances? Explain your reasoning both in terms of distribution of charge on the hollow sphere and the concept of potential, showing internal consistency among the lines of reasoning.

5.17 We have a pair of large capacitor plates oriented and charged as shown in the following diagram. We also have a particle of mass m , carrying charge $+q$ so small that its presence has negligible effect on the charge distribution on the plates. No gravitational or other forces are to be considered. The potential difference between the plates is denoted by ΔV , the spacing between the plates by Δs , and the magnitude of the electrical field strength between the plates by E



- (a) Suppose that our test particle is placed successively at positions A, B, and C. Show, on *separate* force diagrams for the particle, the force acting on the particle at each one of these positions. If the force happens to be zero, say so explicitly. Explain your reasoning.
- (b) If the test particle is displaced from position D to position E, what amount of work must be done on (or taken out of) the system? Explain your reasoning.
- (c) If the test particle is displaced from position F to position G, what amount of work must be done on (or taken out of) the system? Explain your reasoning.

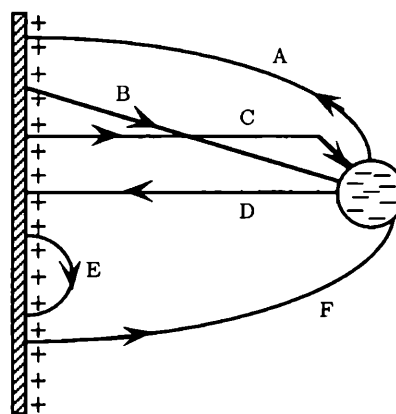
5.18 Two positively charged conducting spheres A and B are located with their centers 10.0 cm apart. The two charges are *unequal*, A carrying 2.5×10^{-8} C and B carrying 1.2×10^{-8} C. Draw a force diagram for each sphere, assuming the presence of fixed supports that keep the spheres from accelerating. Calculate the force exerted by A on B and the force exerted by B on A. Explain your reasoning.

5.19 The charged particles A, B, and C, occupy fixed positions at the vertices of a right triangle, as shown. The charges on the particles are all equal in magnitude. No interactions other than electrostatic are present.

A \ominus B \ominus C \oplus

Draw a force diagram for each particle showing all the forces acting on it. Describe each force in words. Identify the third law pairs. (Show larger forces with longer arrows and smaller forces with shorter arrows.)

5.20 The following diagram shows the region in the neighborhood of a large positively charged conducting plate (extending far beyond the region shown) and a negatively charged conducting sphere. Lines labeled A - F are shown and are said to be field lines for the electric field between the two charged objects.



- Examine each one of the lines, and indicate whether it is a correctly drawn field line. If any line is not correct in some way, explain what is incorrect about it.
- Redraw the diagram to make the pattern of field lines more nearly correct.

5.21 Four particles, each carrying the same magnitude of electrical charge, are located at the corners of a square as shown. Point A is located at the midpoint of one side of the square; point B is located precisely at the center.

 \ominus \oplus

A •

B •

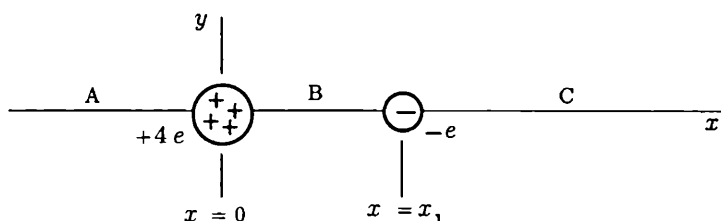
 \oplus \ominus

- Determine qualitatively (numbers are not needed) whether there is a nonzero electrical field at each point (A and B). If the field is not zero, indicate its direction. Show how you came to your conclusions.
- Let us return to the situation at point B (no gravitational effects being present). Suppose you placed a charged particle at B. Would it stay there? Why or why not? (Compare this situation with trying to balance a ball bearing on the very top of a bowling ball. Such points are called positions of "unstable equilibrium.")

5.22 Two charged particles are located along the x -axis, as shown in the following figure. We define region A as that for which $x < 0$, region B as that for which $0 \leq x \leq x_1$, and region C as that for which $x > x_1$.

- The electrical field strength along the x -axis is, of course, zero at both $+\infty$ and $-\infty$. Is the electrical field strength zero at any other point (or points) along the x -axis? Explain your reasoning.

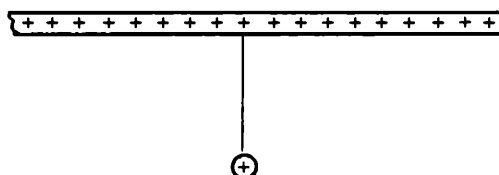
- (b) Identify the precise location of the point at which the field strength is zero. (The calculation does not require an algebraic analysis; it can easily be done mentally.)



- (c) Now consider a more general case: suppose the two particles, located as in the figure, might carry any reasonable quantity of charge of either variety. Are there cases in which there is no point, other than at infinity, where the electrical field is zero along the x -axis? If so, describe them. Are there cases in which the electrical field will be zero at more than one point (other than at infinity) along the x -axis? If so describe them.

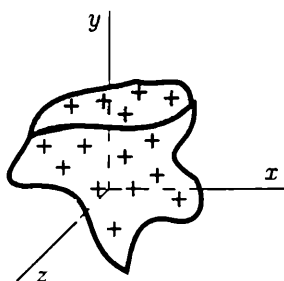
5.23 A positively charged ball is suspended on a nylon (non-conducting) thread from a positively charged metal plate of very large extent, as shown in the diagram. The ball swings back and forth as a pendulum bob. (Both electrical and gravitational effects are present.)

- (a) Draw separate force diagrams for the ball and the thread at some arbitrary angle of deflection from the vertical, and describe each force in words. How will the tension in the string in the presence of electrical charge compare with the tension that would obtain in the presence of gravity alone? Explain your reasoning.



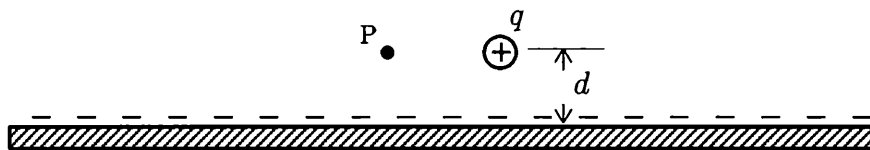
- (b) How will the period of the swing in the presence of electrical charge compare with the period in the presence of gravity alone? Will it be equal to, greater than, or less than the latter? Explain your reasoning.

5.24 Suppose an irregularly shaped three-dimensional region such as that sketched contains $+3.6 \times 10^{-8}$ C of positive charge uniformly distributed throughout the region. There is no charge present outside the region indicated. Suppose we now scale the three-dimensional region down, shrinking each of the x , y , and z dimensions by a factor of 1.85 while keeping all the original charge within the smaller region.



- (a) How would the final charge *density* τ_2 compare with the initial charge density τ_1 ? Calculate the numerical value of the ratio and explain your reasoning. (The term “charge density” refers to amount of charge per unit volume.)
- (b) As we shrink the region down uniformly from all sides, what will happen to the force exerted on a spherical charge of -4.8×10^{-7} C which remains at a fixed position at a distance very large compared with the dimensions of the positively charged region we are shrinking, i.e., will the force on the distant negative charge increase, decrease, or remain essentially unchanged? Explain your reasoning.

5.25 Point P is located at a distance d from a negatively charged metal plate that has dimensions very much larger than d . A small positively charged metal sphere, carrying charge q , is located near P at the same distance d from the plate.



If point P and the metal sphere are both moved up to distance $2d$ from the plate, how will the electrical field strength at the new location of P compare with the field strength at the initial location? Will it be increased, decreased, or essentially the same? Explain your reasoning.

5.26 Recall the phenomena associated (1) with frictional effects, such as rubbing various objects together; (2) with chemical effects, such as those taking place in (voltaic) batteries and in electrolysis; and (3) with effects taking place in so-called “electromagnetic generators.” It is asserted that these are all manifestations of electrical effects involving transfer of electrical charge. In other words, all these seemingly very different phenomena are intimately related.

Suppose you are confronted by a well-educated nonscientist friend who is, very reasonably, skeptical of this assertion, being conscious of the immense disparity among the various phenomena and not believing them to be linked to the same basic concept. How would you proceed to convince him or her that the assertion is indeed correct? Your argument should include appeal to experiments and demonstrations that might be carried out, not just verbal argument. (Your friend is in good company. Faraday felt it necessary to devote an entire, important paper to listing and describing just such demonstrations supporting the relatedness of the phenomena.)

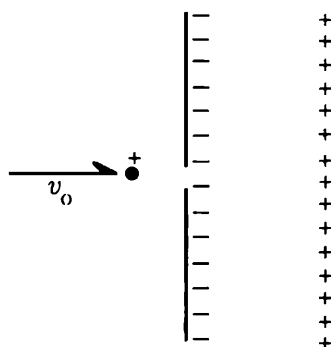
5.27 An electrical engineer files a patent claim on an arrangement of charged electrodes specially shaped so that, in a charge-free cylindrical region the axis of which is parallel to the x -axis, there exists an electrical field with lines of force precisely parallel to the x -axis and with magnitude of field strength increasing linearly according to the equation $E = bx + d$. The equation applies in the region $0 \leq x \leq a$, and b and d are constants.

- (a) Sketch a diagram of the situation that has been described.
- (b) If you were the patent examiner, would you approve or disapprove the claim? Explain your reasoning. (If you have studied Gauss’s law, present your argument by application of this law.)

Note to the student: In the following multiple-choice questions, circle the letters designating those statements that are true or correct. *Any number* of statements, *not* just one, may be correct. You must consider each statement on its merits and not simply stop when you have found one correct statement.

5.28 A positively charged particle carrying 2.0×10^{-8} C enters a region between charged capacitor plates through a hole in one plate, as shown. The potential difference between the plates is 1000 V, and the kinetic energy of the particle as it enters the hole is 1.0×10^{-5} J. (Only electrical effects are to be considered. Gravitational effects and air resistance are to be ignored.)

- (a) The kinetic energy of the particle remains unchanged as it moves toward the right-hand plate.
- (b) The kinetic energy of the particle decreases as it moves toward the right-hand plate.
- (c) The particle has insufficient kinetic energy to reach the right-hand plate, and it “falls back” toward the hole after going part way.
- (d) The particle collides with the right-hand plate and bounces back toward the left-hand one.



- (e) As the particle moves toward the right-hand plate, the potential energy of the particle-capacitor system increases.
- (f) The data given are insufficient to allow calculation of the force acting on the particle when it is between the plates.
- (g) The momentum of the particle is conserved throughout its motion between the plates.
- (h) None of the above.

5.29 A capacitor is first charged by connecting it, by means of wires and a switch, to a battery having a potential difference of 120 V. The switch is then opened, and the capacitor remains charged. What happens when a metal wire is inserted, connecting the charged plates?

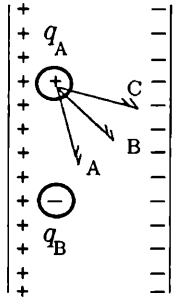
- (a) A spark jumps from one plate to the other.
- (b) The charge on the plates is reversed.
- (c) The charge on both plates becomes zero.
- (d) The potential difference between the plates becomes zero.
- (e) The electrical field strength between the plates is reversed.
- (f) The electrical field strength between the plates drops to zero.
- (g) The kinetic energy of the system is converted to potential energy.

- (h) The capacitance becomes zero.
- (j) None of the above.

5.30 A particle carrying positive charge q_A is subjected to the superposition of two electrical fields—one due to the charged capacitor plates and the other due to the negatively charged particle q_B as shown. The magnitude of charge q_B and its distance from q_A are such that the electrical field strength due to q_B at the position of the positively charged particle is equal in magnitude to the field strength due to the charged capacitor.

Which arrow in the diagram best shows the direction of instantaneous acceleration of the positively charged particle?

- (a) Arrow A.
- (b) Arrow B.
- (c) Arrow C.
- (d) Any of the above, depending on the instantaneous velocity of the positively charged particle in its given location.
- (e) None of the above, since the two fields cancel each other.

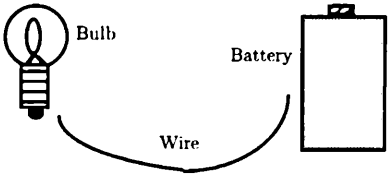


Chapter 6

Direct Current Circuits

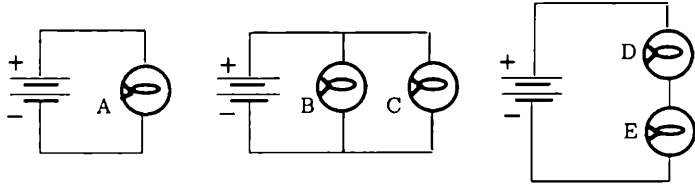
6.1 You are given a flashlight battery, a flashlight bulb, and a single flexible wire as shown.

- (a) Sketch at least two or three different arrangements of these three items that would result in lighting the bulb.
- (b) Sketch at least two or three different arrangements that would not result in lighting the bulb.



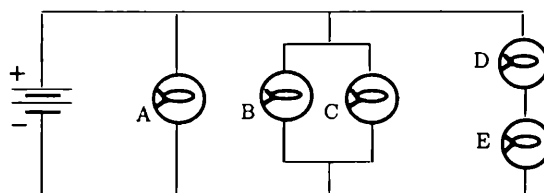
Note to the instructor: Many students have trouble translating a circuit layout on the laboratory table into a conventional circuit diagram and vice versa. Physical as well as pencil-and-paper exercises in such translation, in both directions, are well worth posing for homework and on tests in order to help generate control of this mode of thinking. Complex configurations are not to the point. Simple configurations, such as those illustrated in some of the following problems, are fully adequate to this purpose.

6.2 In these three circuits, identical bulbs, in different combinations, are lighted by connection to identical, ideal batteries. (Ideal batteries have zero internal resistance and suffer no drop in potential difference across their terminals when an external load is connected.)



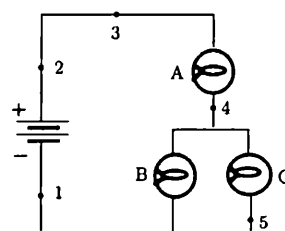
- (a) Rank the five bulbs (A, B, C, D, E) in order of decreasing brightness, indicating equal brightness when such is the case. Explain your reasoning.
- (b) Rank the three circuits in order of increasing current drawn from the battery. Explain your reasoning.

- (c) How does the situation in the following diagram compare with that in part (a)? Is it the same or different? If there are differences, describe them. Be sure to explain your reasoning.



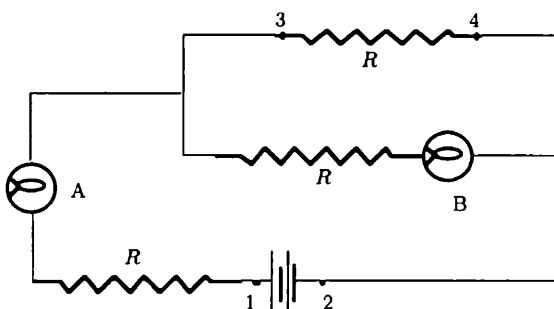
6.3 In the circuit shown in the following diagram, the battery maintains a constant potential difference across its terminals as various changes are made in the circuit containing the three identical bulbs (A, B, and C). You will be asked to predict what will happen as the various changes are made.

- (a) To begin with: How do the brightnesses of the three bulbs compare with each other in the initial condition as sketched? Explain your reasoning.
- (b) Suppose bulb C is removed from its socket. Will the brightnesses of bulbs A and B change? If so how? What will happen to the current at point 3? Explain your reasoning.



- (c) Return to the initial condition in the diagram. Suppose a wire is connected between points 3 and 4 in the circuit. What will happen to the brightness of each bulb? What will happen to the current at point 2? What will happen to the potential difference between points 2 and 4? To the potential difference between points 4 and 5? What will happen to the current at point 5? Explain your reasoning.
- (d) Return to the initial condition. Suppose a wire is connected between points 4 and 5. Answer the same questions as those asked in part (c). Explain your reasoning.
- (e) Return to the initial condition. Suppose a third bulb, D, is added to the circuit by being placed in parallel with B and C. Answer the same questions asked in part (c). Explain your reasoning.
- (f) Return to the initial condition in the diagram. Suppose a wire is connected between points 1 and 5 in the circuit. What will happen to the brightness of each bulb? To the current at point 3? To the potential difference between points 3 and 4? Explain your reasoning.
- (g) Make up a configuration of your own and investigate it by asking questions and making predictions of the variety illustrated above, together with any additional questions you can invent for yourself. One of the best things you can do to strengthen your understanding of simple electric circuits is to obtain or borrow some batteries, bulbs, sockets, and wire and test your own predictions of behavior in various configurations. (The kind of thinking you are asked to do in this question is exactly the kind of thinking that is used in trouble-shooting virtually any electric circuit.)

6.4 The following circuit contains two identical flashlight bulbs A and B and three identical resistors R . The battery maintains a constant potential difference across the circuit regardless of the various changes proposed in the questions. Be sure to explain your reasoning in each case.

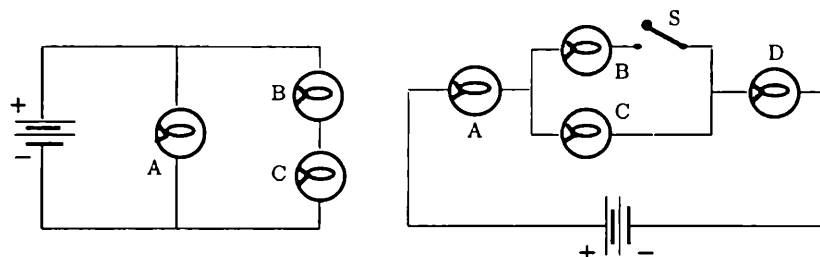


- How do the brightnesses of bulbs A and B compare initially?
- Bulb A is removed. What happens to the brightness of B?
- Bulb A is replaced, and then B is removed. What happens to the brightness of A?
- Bulb B is replaced. A wire is connected from point 1 to point 3. What happens to the brightness of each bulb? What happens to the potential difference between points 3 and 2?
- Return to the initial situation. A wire is connected from point 3 to point 4. What happens to the brightness of each bulb? What happens to the potential difference between points 2 and 3? To the potential difference between points 1 and 3?
- Return to the initial condition. A wire is connected from point 2 to point 4. What happens to the brightness of each bulb and to the potential difference between points 1 and 4?
- Return to the initial condition. A fourth resistor, identical to the other three, is connected between points 3 and 4. What happens to the brightness of each bulb and to the potential difference between points 3 and 4?

Note to the instructor: It is obvious that any number of similar questions can be formed with different (simpler or more complex) configurations. For example, one can construct questions parallel to 6.3 and 6.4 in connection with the slightly simpler circuit that follows (left) or the slightly more complex circuit (right). In connection with the circuit on the right, one might additionally ask about the effect that opening or closing the switch S would have on the brightnesses of the bulbs. One might also add bulbs to an initial configuration.

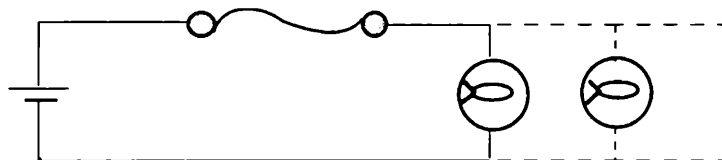
It takes several exercises of this kind to bring a substantial number of students to the point of dealing with them correctly—even students who manage conventional end-of-chapter problems using Ohm's law and Kirchhoff's laws. After giving one or two examples, it is advantageous to challenge students to make up their own

configurations and questions in a contest to outguess the instructor. Those students who accept the challenge rapidly improve their performance and exhibit higher morale as well as satisfaction in achievement.



6.5 In the light of what you have observed happening with series and parallel electric circuits, what do you infer to be the arrangement utilized in your household electric system? Are lights, appliances, etc., connected in series or in parallel or in some combination thereof? Explain your reasoning and your available evidence.

6.6 This circuit consists of a flashlight battery, a flashlight bulb, and a filament of steel wool of grade 0 or 1 (wiggly line) inserted, as sketched, in series with the bulb. The bulb is lighted under these circumstances.



- Suppose you now proceed to add more and more additional bulbs in parallel with the first one as shown by the dashed lines. What do you expect might happen at some point to the filament of steel wool? (Try such a set-up yourself and see. The fact is that the filament burns out.)
- What does the observed burning out of the filament indicate regarding the magnitude of the electric current drawn from the battery as the number of parallel bulbs is increased? Explain your reasoning.
- Explain the analogy between the situation described above and the role played by fuses or circuit breakers in your household electrical system or in your car or radio. What happens in the house as you plug more and more appliances into the same outlet (or into outlets on the same circuit)? Why are fuses or circuit breakers required by law?
- How would you expect the lifetime of a battery used to light one bulb to compare with the lifetime of an identical battery used to light two bulbs in parallel? Explain your reasoning.

6.7 Any conducting object has its own resistance to electric current. Suppose we insert more and more such objects in series across the terminals of a given battery.

- (a) Describe the observations that indicate what happens to the current drawn from the battery as the objects are added. In the light of what happens to the current, what do you infer is happening to the *total* resistance connected across the battery? How does the *total* resistance compare with the resistance of any one of the objects in the system? Explain your reasoning.
- (b) Suppose instead of adding the objects in series, we keep adding them in parallel. Describe observations that indicate what now happens to the total current drawn from the battery. In the light of what happens to the total current, how does the combined effective resistance of the group of objects in parallel compare with the individual resistance of any one of the objects? Explain your reasoning.

6.8 Consider all the experiences you have now had with electric circuits (including experiments, demonstrations, observations made at home, etc.). In all the combined experience at the *macroscopic* level, is there any identifiable evidence concerning the *direction* of flow of either positive or negative charge? (“Yes” or “no” is not an adequate answer. It is necessary to discuss observational evidence.)

6.9 Is it possible to connect two identical batteries to make two bulbs in series light just as brightly as one alone across one battery? If so, sketch and explain the circuit diagram. If not, explain your reasoning.

6.10 Suppose you have lighted a flashlight bulb by connecting it across a battery. While the bulb is lighted, you now connect a wire directly across the terminals of the bulb or its socket. What is observed to happen to the brightness of the bulb? How do you interpret the effect in terms of the concepts of “current” and “resistance”? In terms of our model of what happens in electric circuits, what do you infer happens to the current that passes through the bulb? Does it literally drop to zero or does it level off at some other value? Why is this arrangement unhealthy for the battery? Explain your reasoning.

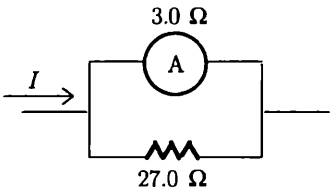
6.11 You have certainly heard the term “short circuit” used in a variety of circumstances. Define this term *operationally*. (This means that you must describe actions and observations of results of these actions. Words or synonyms alone are not adequate.)

6.12 Examine a switch, a socket, and a light bulb if you have never yet done so. It is worth breaking an old light bulb (well wrapped in heavy cloth) to be able to see what is inside, but this should be done with great care.

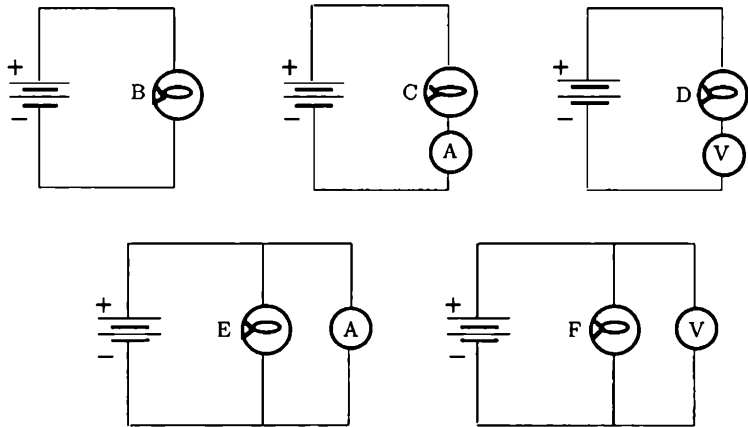
Do switches, sockets, and light bulbs have nonconducting material as well as conducting material present in their structure? If nonconducting material is present, what role does it play? That is, what would happen if it were not there? As part of your explanation, sketch the basic structure of each one of these objects.

6.13 A current of magnitude I divides so as to pass through an ammeter having a resistance of $3.0\ \Omega$ and a resistor of $27.0\ \Omega$.

- (a) What fraction of the current I passes through the meter? Explain your reasoning.
- (b) What would happen to the magnitude of this fraction if the resistance of the resistor were increased? Explain your reasoning.



6.14 In the following circuit diagrams, the circle around the symbol A represents an ammeter; the circle around the symbol V represents a voltmeter. Consider these circuit diagrams in terms of what you have learned about the properties and function of ammeters and voltmeters.



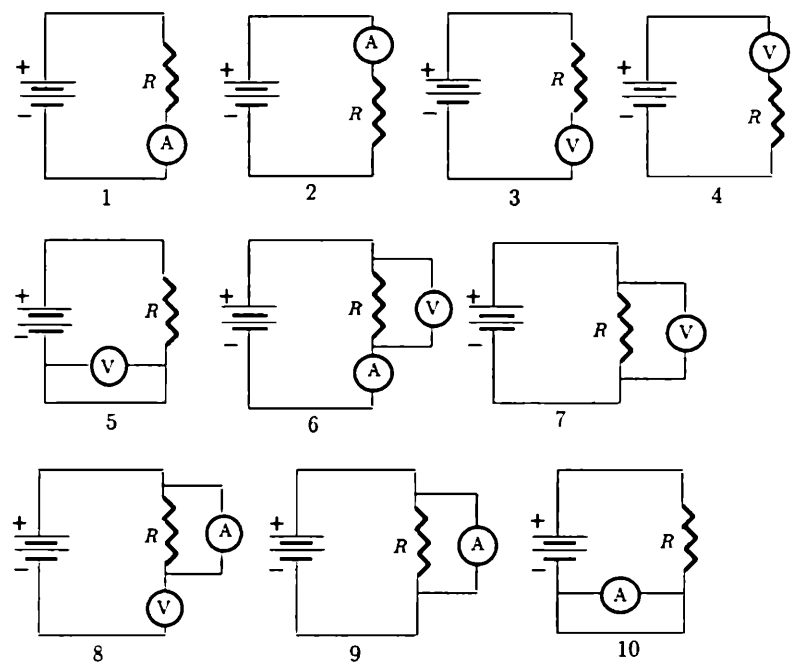
- (a) Compare the brightnesses of the five bulbs (B, C, D, E and F) with the circuits wired as indicated. Explain your reasoning.
- (b) Explain why circuit E is not healthy for either the battery or the ammeter.

6.15 Suppose you are given two boxes that conceal what is inside, although each has two wires coming out of the interior. You have at your disposal some batteries, wire, and identical flashlight bulbs. Describe how you might go about determining which of the two boxes has the lower electrical resistance or whether the two are identical. Show the diagrams you would employ and explain your reasoning.

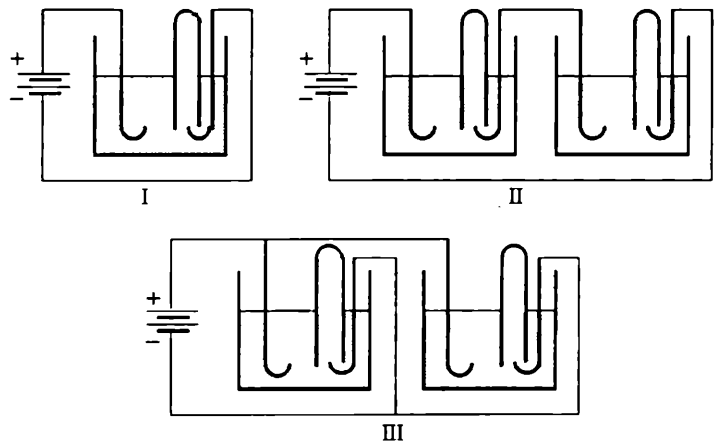
6.16 The next illustration shows ten circuits containing an unknown resistance R and various arrangements of ammeters and voltmeters. Explain your reasoning in answering each of the following questions.

- (a) With which circuit (or circuits) would the arrangement allow measurement of the value of the resistance R if the potential difference at the battery terminals is unknown?
- (b) With which circuit (or circuits) would the arrangement allow measurement of the value of the resistance R if the potential difference at the battery terminals is known?

- (c) In which circuit (or circuits) would the voltmeter read zero?
- (d) In which circuit (or circuits) would the ammeter read zero?
- (e) Which circuit (or circuits) would burn out the ammeter?
- (f) Which circuit (or circuits) would burn out the voltmeter?
- (g) Which circuit (or circuits) would allow determination of the power dissipated in R if the value of R is known?



6.17 The following electrolysis circuits, I, II, and III, contain identical electrolytic cells connected to identical batteries. The oxygen is allowed to escape to the air while



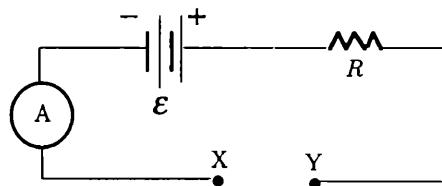
the hydrogen is collected in the inverted test tubes, which are initially completely filled with the electrolytic solution. All three circuits are run for the same length of time at room temperature and pressure, and hydrogen is collected.

- Examine the circuits and indicate which involve series and which involve parallel arrangements of electrolytic cells.
- How would you expect the volumes of hydrogen in the various test tubes to compare with each other? Label the test tubes, and give their comparative rank, being sure to indicate equality of volumes when that is the case. Explain your reasoning.
- How would the resistances of arrangements II and III compare with the resistance of I?

6.18 Two different light bulbs, A and B, are connected in parallel across the same battery. Bulb A lights more brightly than bulb B. In what property do the two bulbs differ? How do they differ? Explain your reasoning.

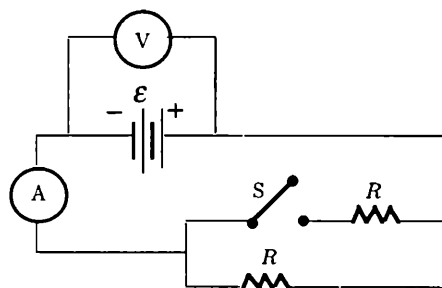
6.19 An *open* circuit contains a resistance R , an ammeter A, and a battery having emf ϵ and internal resistance r , as sketched in the following diagram. Explain your reasoning in answering each of the questions.

- Under these circumstances, what is the potential difference between points X and Y?
- What is the potential difference across the resistor R ?



- What is the potential difference across the battery terminals?
- Does the potential difference across the battery terminals change if a wire is now connected between X and Y? If so, how does it change?

6.20 Consider the following circuit in which the battery has an emf ϵ and internal resistance r . A voltmeter V and ammeter A are placed as shown. When the switch S is closed, how do the voltmeter and ammeter readings compare with the readings that obtained before S was closed? That is, do they increase, decrease, or remain unchanged? Explain your reasoning.

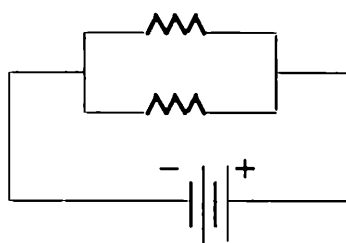


6.21 A wire of one metal has twice the resistivity, twice the diameter, and twice the length of a wire of another metal.

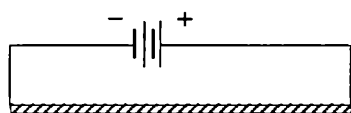
What is ratio of the resistances of the two wires? Give a numerical value and explain your reasoning.

6.22 Two resistors are connected in parallel across a battery with negligible internal resistance, as shown. One of the resistors carries a current of 1.00 A while the other carries a current of 2.00 A.

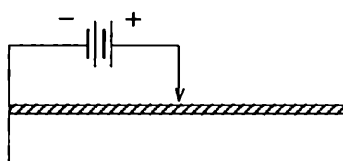
Approximately what will be the current in the circuit if the two resistors are now connected in series instead of in parallel? Explain your reasoning.



6.23 A long, uniform resistive wire has a resistance such that, when it is connected to a battery as shown in circuit (a), the current in the circuit is 1.00 A. The circuit is now altered in such a way that the moveable contact (indicated by the arrow) is positioned at the center of the resistive wire rather than at the right-hand end as in (b), and the right-hand end of the wire is connected back to the left-hand end. Assume that the battery has negligible internal resistance.



(a)



(b)

What will be the current through the battery in circuit (b)? [You might find it helpful to redraw circuit (b) in a way that makes the pattern of resistances more familiar.] Explain your reasoning.

6.24 A circuit contains three identical bulbs lighted by connection to a single battery. When a wire is connected across the terminals of one of the bulbs, it goes out and so does another one of the bulbs. The remaining bulb becomes brighter.

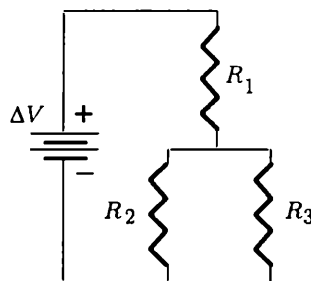
- Sketch the circuit in which the bulbs must have been arranged. Explain your reasoning.
- What would happen to the brightnesses of the other two bulbs if the wire were connected across the terminals of the bulb that does not go out in part (a)?

6.25 Suppose we have two wires X and Y of different metals. The wires have the same resistance at room temperature, but the resistance of X increases more rapidly with increasing temperature than does that of Y. (If their temperatures are equal, the wires lose heat to the surroundings at the same rate.)

- If the wires are connected in parallel across a battery, which wire will have the higher temperature when the system reaches equilibrium? Explain your reasoning.
- If the wires are connected in series across a battery, which wire will have the higher temperature when the system reaches equilibrium? Explain your reasoning.

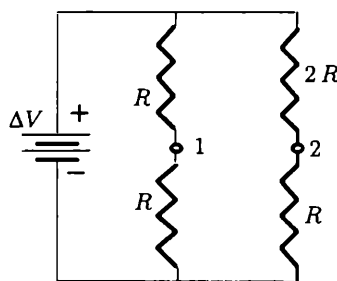
6.26 Three resistors are connected to a battery as shown. Let us suppose that the battery maintains a constant potential difference ΔV across its terminals and that the resistances differ in magnitude so that $R_1 > R_2 > R_3$.

- Rank the current drawn from the battery and the currents in each of the three resistors in order of increasing magnitude. Explain your reasoning.
- Rank the potential difference across the battery terminals and the potential differences across each of the three resistors in order of increasing magnitude. Explain your reasoning.
- Describe in words the circumstances under which the potential difference across R_1 would be equal to that across R_2 .



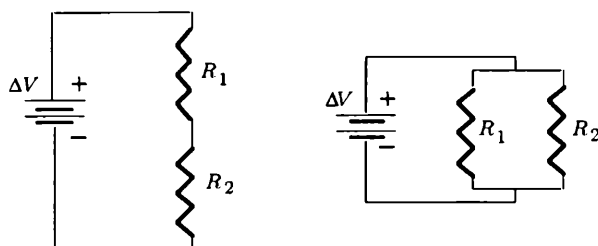
6.27 Four resistors, connected to a battery as shown in the following diagram, have their relative values indicated. The battery maintains a constant potential difference ΔV across its terminals.

- If you connected a voltmeter between points 1 and 2, would it measure a finite or a zero potential difference? Explain your reasoning. If the potential difference is not zero, which point is more positive? (The upper terminal of the battery is the positive terminal.) Explain your reasoning.
- Suppose a wire is now connected between points 1 and 2. (This makes the potential difference between points 1 and 2 zero.) Will there now be a current in this wire? If so, in what direction (assuming conventional positive current)? Explain your reasoning.



- Suppose the upper right-hand resistor in the diagram had the same resistance R as the other three resistors instead of the value $2R$. Answer the same questions as in parts (a) and (b).

6.28 Consider these two circuits with identical batteries. In each case resistance R_1 is greater than R_2 .

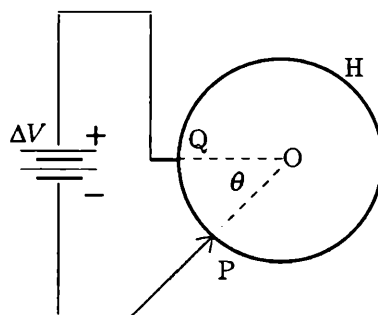


- In each circuit, which resistance dissipates the larger amount of power? Explain your reasoning.

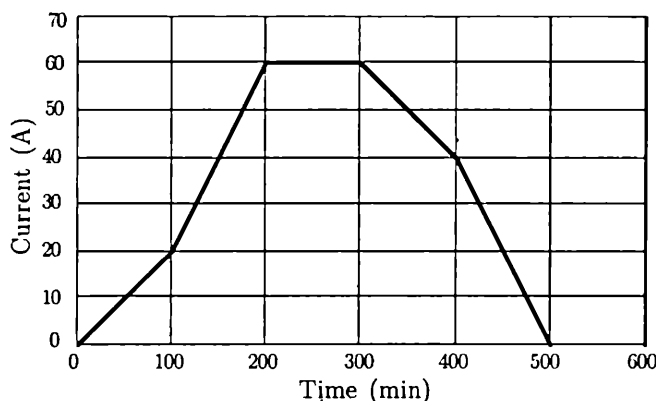
- (b) Explain how it comes about that the same resistance does not dissipate the larger amount of power in each of the circuits.
- (c) In discussions of power P dissipated in a resistor obeying Ohm's law, it is shown that two relations are possible: (1) $P = I^2 R$ and (2) $P = (\Delta V)^2 / R$. Are they both applicable to any situation that arises? Why or why not? What relevance does this question have to your answers in parts (a) and (b)?

6.29 H is a resistive metal hoop to which a battery is connected at points P and Q. Contact Q is fixed while contact P can be moved at will around the hoop, varying the angle θ from 0 to 2π .

- (a) What is the value of the resistance across the battery when P coincides with Q? Explain your reasoning.
- (b) Does the resistance across the battery change as contact P is moved counter-clockwise starting at Q? If so, does it increase or decrease? Explain your reasoning.
- (c) Is there a position of contact P at which the resistance across the battery is greatest? If so, where? Explain your reasoning.



6.30 The graph shows a record of current in amperes versus time in minutes. The current was drawn from a 120 V direct current source to operate some direct current motors.



If the cost of electric power under these circumstances is 5.0 cents per kilowatt-hour, calculate the total cost in dollars of the electrical energy supplied. Show and explain all steps of your calculation and explain your reasoning.

6.31 A wire is connected to a battery. Starting at room temperature, the temperature of the wire increases until it levels off at some final value, higher than the initial. The wire is still connected to the battery and electric current is still flowing.

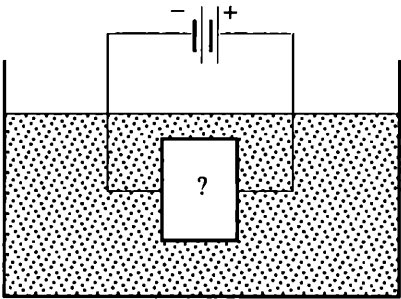
- (a) How do you account for the fact that the temperature of the wire does not continue to increase indefinitely?

- (b) Explain the role of the resistance of the wire in determining the temperature that is attained.
- (c) Under what circumstances does the wire melt or “burn out”?

Note to the student: In the following multiple choice questions, circle the letter or letters designating those statements that are correct. *Any number* of statements may be correct, *not* just one, and you must examine each statement on its merits.

6.32 Suppose you are given an ideal battery that maintains a constant potential difference ΔV across its terminals regardless of which of the following resistive loads is connected to it. You are also given three identical resistors, each with a resistance of R ohms. You wish to heat water in a beaker with a setup such as that sketched in the following diagram.

- (a) Which of the following combinations of the resistors would you put into the box with the question mark to produce the most rapid heating of the water?
- (b) Explain your reasoning, and, if you choose (G), indicate what other combination you would introduce and why.



- (A)

(B)

(C)
- (D)

(E)

(F)

(G) NONE OF THE ABOVE.

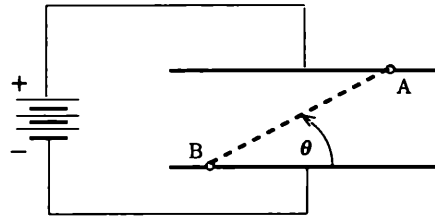
Note to the instructor: One can alter question 6.32 to demand more courage of conviction by leaving out configuration C and making the correct answer G, asking students to sketch the correct combination if G is chosen.

6.33 A capacitor is charged by connecting it to a battery and is then disconnected from the battery by opening a switch. When a wire is inserted, connecting the charged plates:

- (A) a spark jumps between the plates.
- (B) the charge on both plates becomes zero.

- (C) the charge on the plates is reversed.
- (D) the potential difference between the plates becomes zero.
- (E) the kinetic energy of the system is converted to potential energy.
- (F) the electrical field strength between the plates becomes zero.
- (G) the electrical field strength between the plates is reversed in direction.
- (H) None of the above.

6.34 A capacitor is connected to a 120 V battery as shown in the diagram. A very high resistance wire is then inserted, connecting points A and B on the plates.



The potential difference (in volts) between the ends of the wire will then be

- | | |
|-------------------------|-------------------------|
| (A) $120 \cos \theta$ | (D) $120 \sin \theta$ |
| (B) 120 | (E) $120 / \cos \theta$ |
| (C) $120 / \sin \theta$ | (F) $120 \tan \theta$ |

Chapter 7

Electromagnetism

Note to the instructor: Most students need drill exercises, both in homework and on tests, with the mnemonics concerning the basic electromagnetic phenomena. Few texts provide enough of these exercises to help fix the rules in student memories even within the immediate time interval of an ongoing course. Illustrative exercises, ranging from exceedingly simple to slightly more sophisticated, are given here. Such questions could easily be programmed on computers and could provide most of the needed drill. Variations on the themes illustrated are readily generated.

7.1 The diagram shows the cross section of a wire carrying conventional positive electric current away from us into the plane of the paper.

- (a) By means of an arrow on the diagram, show the direction in which a compass would point if placed at location A and describe the rule you use to help remember this effect.
- (b) Show the direction in which the compass would point at two or three other locations of your own choosing.



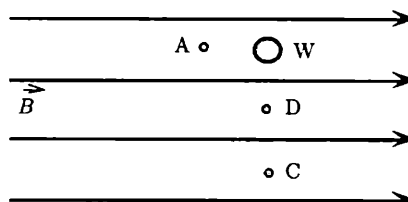
•A

Note to the instructor: Variations on the preceding question might include (1) giving the direction of the B-field and asking for the direction of conventional current, (2) placing the current-carrying wire in the plane of the paper and asking for the B-field direction in or out of the plane, (3) given a B-field direction in or out of the plane, asking for the direction of current in the wire. The orientation of the wire should be varied in the plane of the paper; it should not be restricted to horizontal or vertical. Similar variations in the geometry of an arrangement can be invoked in many of the questions that follow.

7.2 A uniform B-field is directed toward the right in the plane of the paper, as shown. A wire W, lying perpendicular to the plane of the paper and seen in cross section, carries electric current. The resultant magnetic field at point D is zero.

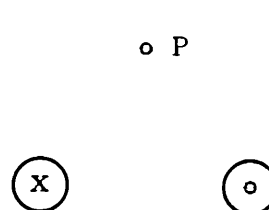
- (a) What must be the direction of current in the wire (into or out of the plane of the paper)? Explain your reasoning.
- (b) Point A lies the same distance from the center of the wire as point D. Choosing a length for the vector representing the horizontal B-field, construct a vector diagram showing the resultant vector at point A. Explain your reasoning.

- (c) Point C lies twice the distance from the center of the wire as point D. Construct a vector diagram showing the resultant vector at point C. Explain your reasoning.



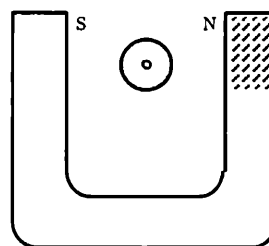
7.3 Two wires lie perpendicular to the plane of the paper and carry equal electric currents in the directions shown. Point P is equidistant from the two wires.

- (a) Construct a vector diagram showing the resultant \vec{B} vector at point P. Explain your reasoning.
- (b) Suppose a third wire carrying equal current into the plane of the paper were located at P. What would be the direction of the force on this wire? Explain your reasoning.



7.4 The diagram shows a horseshoe magnet with its north and south poles marked. Between the poles is the cross section of a wire, oriented perpendicular to the plane of the paper and carrying conventional positive electric current toward us out of the plane of the paper.

- (a) Add to the diagram a few magnetic field lines showing the direction of the B-field (due to the magnet) in the neighborhood of the wire.
- (b) Show the direction of the force acting on the wire under these circumstances and describe the rule you use to help remember the experimental facts of the interaction involved.

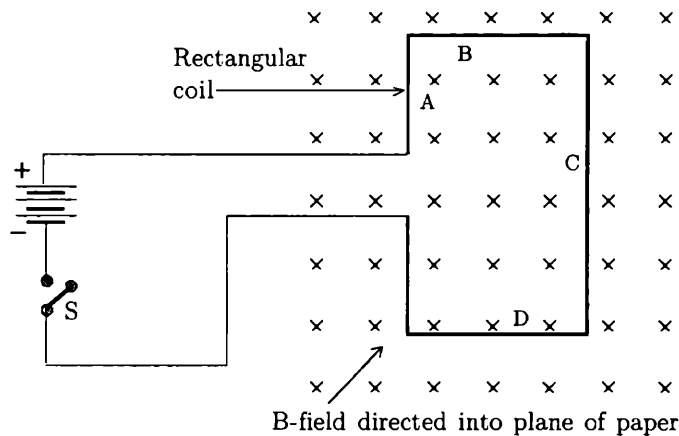


- (c) Does the magnet experience a force under these circumstances? If so, in what direction? If not, why not?
- (d) Suppose the wire is rotated so that it lies in the plane of the paper, with the current directed from right to left. What would be the direction of force on the wire?
- (e) Suppose you are told that a wire carrying current into the plane of the paper experiences an upward force. What do you infer must be the direction of the B-field to which the wire is subjected?

7.5 The following figure shows a rectangular coil of wire connected to a battery through a switch S. The polarity of the battery is indicated. The coil is located in a uniform B-field directed into the plane of the paper as indicated, and the plane of the coil is in the plane of the paper.

- (a) Show the direction of conventional positive current in the coil when switch S is closed, indicating how you arrive your conclusion.

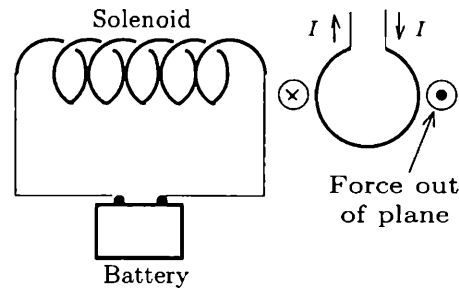
- (b) Show by means of arrows the direction of the force acting on the coil at each of the positions A, B, C, and D when the switch S is closed. If the force is zero at any point, say so explicitly. (Show a force directed into the plane of the paper by the symbol \times and a force directed out of the plane of the paper by the symbol \bullet if needed.)



- (c) In this orientation relative to the B-field, does the coil tend to rotate or does it tend *not* to rotate? Explain your reasoning. If the coil does tend to rotate, indicate the axis and the direction of rotation. If the coil does not tend to rotate in the given orientation, describe the overall effect of the forces to which it is being subjected. If the coil does not tend to rotate in this orientation, describe an orientation in which it *would* tend to rotate.
- (d) Answer the questions in part (c) for the case in which the current in the system is reversed.

Note to the instructor: An obvious variation on Problem 7.5 is to ask essentially the same sequence of questions for the case in which the B-field is parallel to the plane of the coil.

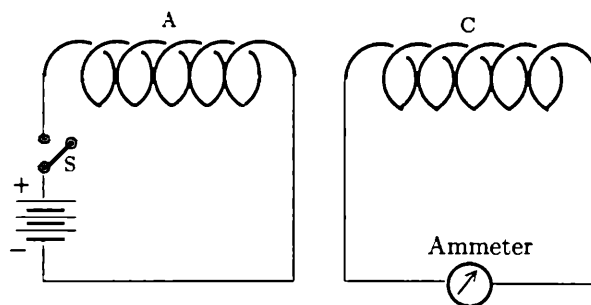
7.6 The diagram shows a solenoid connected to a battery; the axis of the solenoid is parallel to the plane of the paper, and the magnetic field at the end of the solenoid is therefore also parallel to the plane of the paper. Located in the magnetic field of the solenoid is a suspended current-carrying coil, the plane of which is parallel to the magnetic field.



The direction of conventional positive current I in the coil is shown, as are the directions of the forces (out of the paper and into the paper) on either side of the coil, defining the direction of torque on the coil.

From the information given, deduce the polarity of the battery, explaining your steps of reasoning as you go along. Label the polarity on the diagram.

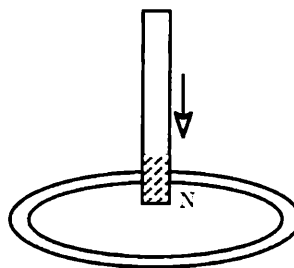
7.7 Two solenoids A and C are sufficiently close together that the magnetic field formed in A, in the presence of electric current, also penetrates into C. (The effect can be greatly strengthened by passing a soft iron rod through both solenoids.)



- We start with switch S closed so that a steady current is present in solenoid A. Establish the direction of the B-field penetrating into solenoid C and show its direction by means of an arrow labeled "B-field." Explain how you arrive at your result. If the B-field is zero, say so explicitly and explain your reasoning.
- While the switch is closed and the steady current is present in A, what is the direction of induced current in C? Show by means of a labeled arrow and explain how you arrived at your conclusion. If the current is zero, say so explicitly and explain your reasoning.
- The switch is now opened, and the current in A drops to zero. Describe what, if anything, happens in C, showing the direction of any possibly induced current. Explain how you arrived at your conclusion. What is the situation in C after the current in A has dropped to zero? Explain your reasoning.

7.8 The north pole of a magnet is thrust downward into a horizontally oriented copper ring as shown.

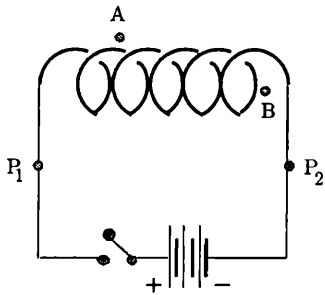
- Does the ring experience a force under these circumstances? If so, in what direction? Explain your reasoning.
- Does the magnet experience a force? If so, in what direction? Explain your reasoning.



7.9 A solenoid is connected to a battery as shown. Switch S is now closed, and there is a current in the circuit.

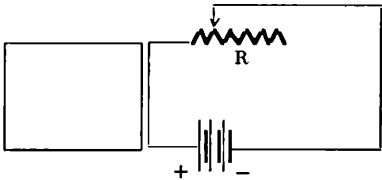
- By means of appropriately placed and labeled arrows, show the direction of conventional positive current I in the circuit.
- Suppose compasses are placed at points A and B near the solenoid. Show by means of labeled arrows the direction in which a compass would point at each location, and explain how you arrive at your conclusions.

- (c) Suppose an unmagnetized iron bar is placed near point B. Will the bar experience a force? If so, in what direction relative to the plane of the paper? If not, why not?
- (d) Suppose a resistor is now connected between points P₁ and P₂ with switch S still closed. What will happen to the strength of the B-field at point B? Will it increase, decrease, or remain unchanged?

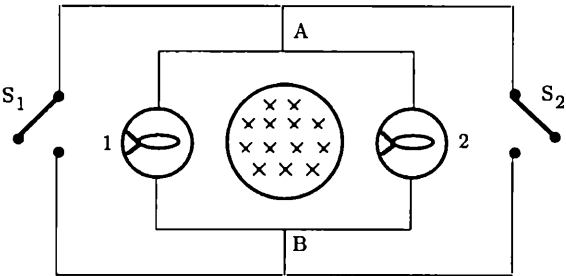


7.10 The diagram shows a resistive, current-carrying circuit on the right and a closed wire loop on the left. Suppose the contact with the resistor R is moved to the right in the right-hand circuit.

Will current be induced in the wire loop in the left-hand part of the diagram? If so, in what direction? If not, why not? Explain all steps of your reasoning.



7.11 Consider the following diagram. A solenoid produces a B-field directed into the plane of the paper. The current in the solenoid is increasing at a uniform rate, causing the total flux directed into the plane of the paper to be increasing at a uniform rate. Two identical flashlight bulbs, 1 and 2, having resistance denoted by r , are connected in a circuit surrounding the solenoid as shown and light up while the flux increase is taking place. Switches S₁ and S₂ are initially open.

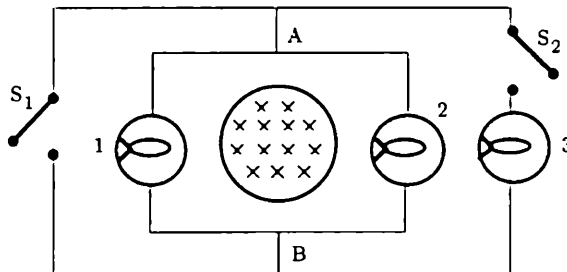


- (a) What is the direction (clockwise or counterclockwise) of the emf ϵ induced in the circuit containing the bulbs? Explain how you arrive at your answer.
- (b) In terms of the quantities ϵ and r , what is the current in the circuit containing the bulbs while the current in the solenoid is increasing? How do the brightnesses of the bulbs compare under these circumstances? Explain your reasoning.
- (c) What happens to the brightness of bulb 2 if bulb 1 is removed from its socket while the current in the solenoid is still increasing? Explain your reasoning.
- (d) Suppose that with both bulbs in their sockets, switch S₁ is now closed while S₂ is left open. What happens to the brightness of each bulb while the current in

the solenoid is increasing? (Explain your answer in terms of what happens to the combined resistance on the left-hand side of the solenoid.) In terms of the quantities ε and r , what is now the current in bulb 2? The current in bulb 1?

- (e) Suppose that switch S_1 is opened and switch S_2 is closed. Answer the same questions asked in part (d).
- (f) Note that, in both parts (d) and (e), we are introducing a wire between points A and B and are thus short-circuiting *both* bulbs regardless of which switch we close. In ordinary batteries-and-bulbs circuits previously encountered, a short across any combination of bulbs would have caused all the short-circuited bulbs to go out. In the system now under consideration, this is not the case. What happens to the brightness of one of the seemingly short-circuited bulbs is determined by the *direction* taken by the short-circuiting wire. Explain in your own words how this profound difference between the two types of circuit arises. (We sometimes speak of the situation involving the solenoid and the induced emf as constituting a “multiply connected” as opposed to a “simply connected” region.)

7.12 Consider the same situation described in the preceding question except that a third identical bulb (3) is introduced in the manner shown, and the current in the solenoid is decreasing at a uniform rate while the B-field is still directed into the plane of the paper.

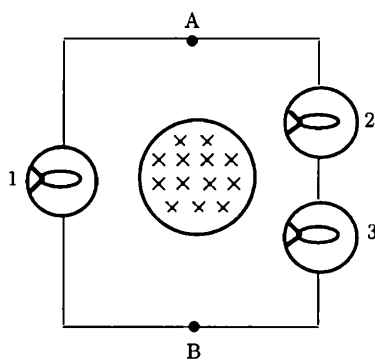


- (a) What is the direction of the induced emf ε in the inner circuit? Explain your reasoning.
- (b) In terms of the quantities ε and r , what is the current in bulbs 1 and 2 while the current in the solenoid is decreasing and both switches are open?
- (c) Suppose that switch S_2 is now closed while S_1 is left open. What will happen to the brightnesses of bulbs 1 and 2 compared to their initial levels? How will the brightnesses of 1, 2, and 3 compare with each other after S_2 is closed? Explain your reasoning.
- (d) What is the total resistance of the circuit before S_2 is closed? What is the total resistance after S_2 is closed? What will be the current in bulb 1 after S_2 is closed? (Answer: $2\varepsilon/3r$) What will be the current in bulb 2 after S_2 is closed? (Answer: $\varepsilon/3r$) Explain your reasoning. Are your results consistent with your qualitative predictions regarding comparative brightnesses in part (c)?

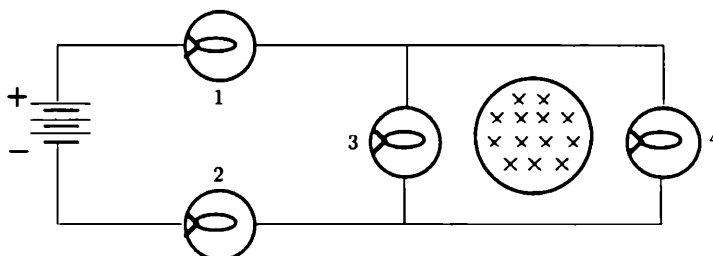
- (e) Answering the same questions as in parts (c) and (d), discuss what will happen if switch S_1 is closed while S_2 is left open.
- (f) What will happen if both switches are closed? How will current and brightnesses compare with the situation that obtains when both switches are open? Explain your reasoning.

7.13 A long solenoid lies with its axis perpendicular to the plane of the paper. A linearly changing current in the solenoid induces a clockwise emf. A wire of negligible resistance connects three identical flashlight bulbs in the manner shown. The bulbs light when the emf is induced.

- (a) How do the brightnesses of the bulbs compare with each other under the circumstances illustrated in the diagram? Explain your reasoning.
- (b) What will happen to the brightness of each bulb if bulb 2 is removed from its socket? Explain your reasoning.
- (c) Restore bulb 2 to its socket. What will happen to the brightness of each bulb if a wire shorts points A and B while connected so that it lies around the right-hand side of the solenoid? How will the brightnesses now compare with those that obtained before the shorting? Explain your reasoning.
- (d) Answer the same question as part (c) if the wire shorts points A and B while lying around the left-hand side of the solenoid. Explain your reasoning.
- (e) Answer the same question as part (c) if the wire shorts points A and B by being connected symmetrically through the center of the solenoid. Explain your reasoning.
- (f) Suppose that two voltmeters are connected between points A and B. One voltmeter (call it V_1) is connected so that it and its wires make a circuit around to the left of bulb 1, while the other meter (call it V_{23}) is connected so that it and its wires make a circuit around to the right of bulbs 2 and 3. Note that both voltmeters are connected between the same two points A and B. In ordinary batteries-and-bulbs circuits, we found that voltmeters connected between the same two points show the same readings no matter what the geometry of the arrangement. Will voltmeters V_1 and V_{23} in this arrangement show the same readings? Why or why not? Explain, as though you were helping a fellow student see what is involved, why it is necessary to be very careful about interpreting voltmeter readings in circumstances that involve induced emf's and multiply connected regions. [If you are interested in pursuing further the issue of the meaning of voltmeter readings, read the article by R. H. Romer "What Do 'Voltmeters' Measure? Faraday's Law in a Multiply Connected Region," *American Journal of Physics*, 50, 1089 (1982).]



7.14 The following circuit consists of four identical flashlight bulbs, each with resistance r , and an ideal battery (negligible internal resistance relative to the bulbs) that maintains a constant potential difference ΔV across its terminals. A solenoid oriented perpendicular to the plane of the paper and producing a B-field directed into the plane of the paper occupies the position shown.

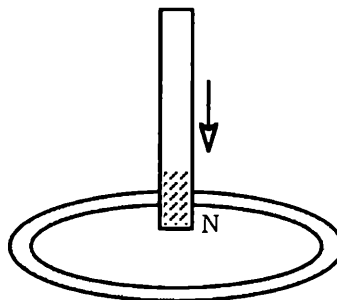


- Initially, with the current in the solenoid constant and the B-field unchanging, find the current in each of the four bulbs in terms of ΔV and r . Explain your solution step by step. How do the brightnesses of the bulbs compare with each other?
- Suppose we now either increase or decrease the current in the solenoid linearly with time. What direction of induced emf ε (clockwise or counterclockwise) do we need to produce around the solenoid to make bulb 3 go out (i.e., to carry zero current)? Explain your reasoning.
- What emf ε (expressed in terms of the known potential difference ΔV) must be induced to produce zero net current in bulb 3? Show and explain all steps of your reasoning. [Hint: Calculate the effective resistance lying to the left of the solenoid and the total resistance in the loop around the solenoid; then find the induced current in the loop and the induced current through bulb 3.] (Answer: $\varepsilon = \Delta V/2$)
- What is the current in each of the bulbs 1, 2, and 4 under the circumstances obtaining in part (c)? Show and explain all steps of your reasoning. (Answer: $\Delta V/2r$)
- For the situation in part (c), what will be the current in each remaining bulb if bulb 4 is removed from its socket? Explain your reasoning.
- For the situation in part (c), what will be the current in each bulb if a wire shorts bulb 4 by being connected around the right-hand side of the solenoid? Show your reasoning. Answer the same question for the case in which a wire shorts bulb 3 by being connected around the left-hand side of the solenoid.
- What magnitude and direction of emf would cause bulb 4 to go out rather than bulb 3? Show and explain all steps of your reasoning.

Note to the student: In the following questions, circle the letter or letters designating statements you believe to be correct. *Any number* of statements may be correct, *not* necessarily just one. You must examine each statement on its merits.

7.15 If the north pole of a magnet is thrust downward into a horizontally oriented copper ring as shown in the following diagram, the ring will experience

- (A) a downward force.
- (B) an upward force.
- (C) a sidewise force.
- (D) zero force.
- (E) a clockwise torque as seen from above.
- (F) a counterclockwise torque as seen from above.
- (G) None of the above.



7.16 The diagram shows a ring of copper with its plane perpendicular to the horizontally oriented axis of the nearby cylindrical magnet. A current will be induced in the ring if

- (A) the magnet is moved toward the ring.
- (B) the ring is moved away from the magnet.
- (C) the ring is rotated about any of its diameters.



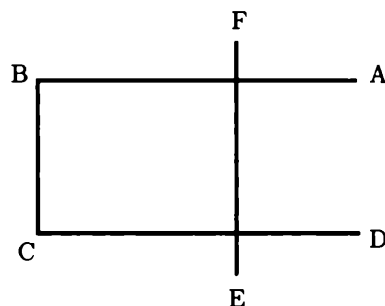
- (D) the magnet is spun around its horizontal axis (the line through its length.)
- (E) the magnet is rotated around a vertical axis through its center.
- (F) the magnet is rotated around a horizontal axis perpendicular to the plane of the paper.
- (G) the magnet is moved up or down.
- (H) the ring is rotated around its center in the plane in which it lies.
- (I) None of the above.

7.17 A small bar magnet lies in a region of nonuniform magnetic field. Under such circumstances, depending on how it is oriented relative to the field direction, the magnet might experience

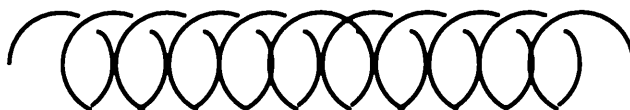
- (A) a torque but zero net force.
- (B) a net force but zero torque.
- (C) both a torque and a net force.
- (D) neither a torque nor a net force.
- (E) translational motion if it is free to move.

7.18 In the figure, we are looking down on a bare wire ABCD lying on a table top. Lying across ABCD and making good electrical contact with it is another bare wire EF which is free to move. A B-field is directed down through the rectangle of wires. If the strength of the B-field is decreased, the wire EF will

- (A) remain stationary.
- (B) slide to the right.
- (C) slide to the left.
- (D) rotate clockwise.
- (E) tend to move upward out of the plane of the paper breaking contact with ABCD.
- (F) tend to move downward into the plane of the paper, pressing against ABCD.
- (G) None of the above.



7.19 The helical coil of wire carries an electric current. As a result of the current, the coil



- (A) tends to unwind.
- (B) tends to become shorter.
- (C) tends to become longer.
- (D) has no tendency to alter its shape.
- (E) behaves in a way that cannot be predicted since the effect depends on the direction of current.

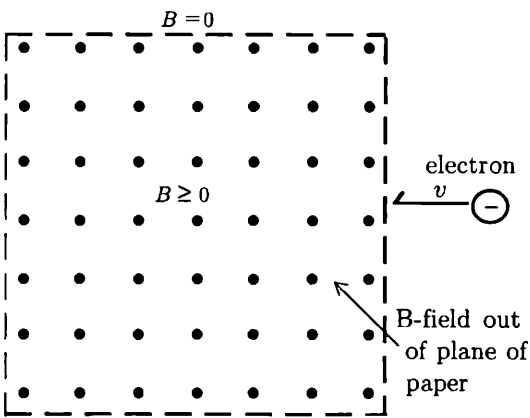
Chapter 8

Particle Trajectories in E- and B-Fields

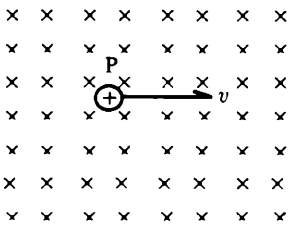
8.1 In an evacuated space, an electron moves with velocity v toward the left in the plane defined by the paper as shown. The electron enters a region of uniform B-field directed out of the plane of the paper. The magnetic field strength B drops very sharply to zero outside the region defined by the dashed lines. Note that, depending on the strength of the field, the electron might exit into the field-free region at various points; be sure to take this into account when sketching trajectories.

Sketch possible trajectories of the electron for each of the following four B-field conditions within the region defined by the dashed lines and label each trajectory with the corresponding number:

- (a) $B = 0$ throughout the region.
- (b) Weak, but not zero, B-field.
- (c) Moderately strong B-field.
- (d) Very strong B-field.



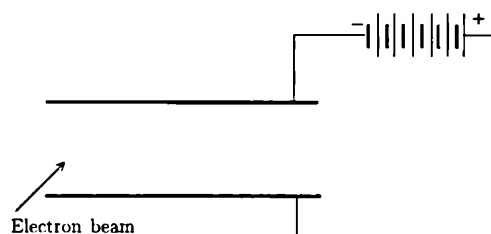
8.2 A uniform B-field is directed into the plane of the paper in an evacuated space. A hydrogen ion (H^+) and an alpha particle (He^{2+}) both have the same velocity v (in the plane defined by the paper) at point P. An alpha particle has four times the mass of a hydrogen ion. The field is so strong that in the absence of collisions, both particles execute complete circular trajectories within the region of the field.



Deduce the relative size of the circles and sketch them to scale on the diagram, showing the direction of travel in each case. Explain your reasoning.

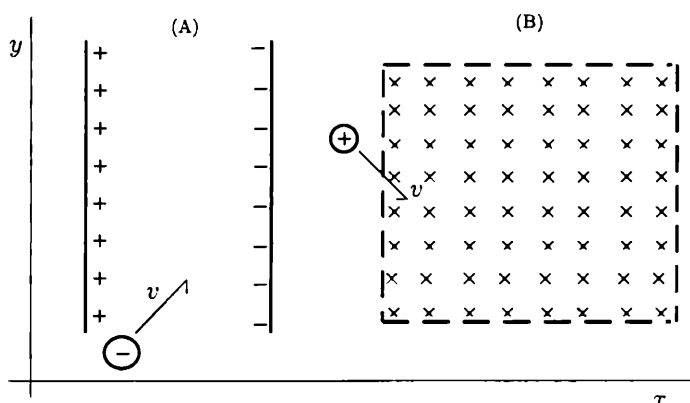
8.3 A beam of electrons in a highly evacuated space enters an E-field between capacitor plates in the direction shown in the diagram.

Sketch a possible electron trajectory for each of the following cases of different E-field strength, remembering that gravitational effects are unobservably small. Sketch each trajectory neatly and carefully—not sloppily and carelessly—and label each one with the appropriate letter. (If the beam exits from between the plates, be sure to show the character of the trajectory after the exit.)



- (a) $E = 0$, i.e., plates uncharged.
- (b) Weak E-field, small potential difference across the plates.
- (c) Moderate E-field.
- (d) Very strong E-field.

8.4 In part (A) of the diagram, a negatively charged particle moving at velocity v in an evacuated space enters a region between capacitor plates as shown. In part (B), a positively charged particle enters a sharply defined region of uniform B-field directed into the plane of the paper.



Examine the following statements. Indicate whether each is true or false and explain the reasoning behind your conclusions.

- (a) The negatively charged particle in (A), on entering the electrical field between the capacitor plates, moves without change in the y -component of its velocity.
- (b) The negatively charged particle in (A), on entering the electrical field between the capacitor plates, moves without change in the x -component of its velocity.
- (c) The positively charged particle in (B), on entering the region of magnetic field, moves without change in the x -component of its velocity.

- (d)

The positively charged particle in (B), on entering the region of magnetic field, moves without change in the y -component of its velocity.
- (e)

Neither particle changes its speed as it continues along its trajectory.

8.5 In the following diagram, an electron, at the position indicated, moves with velocity v in the direction shown. The space is highly evacuated, and a uniform magnetic field of strength B is directed out of the plane of the paper. B drops sharply to zero immediately outside the region shown. Note that the electron may exit the region of the B-field; be sure to take this possibility into account when sketching trajectories.

- (a)

Sketch a possible trajectory of the electron, showing the direction of travel. Explain how we know that the trajectory within the B-field is circular.
- (b)

Explaining all steps, derive an expression for the radius R of the circular part of the trajectory in terms of known quantities such as the mass and charge of the electron, the velocity v , and the magnitude of the field strength B .
- (c)

Interpret the expression you have obtained for R by adding to the figure:

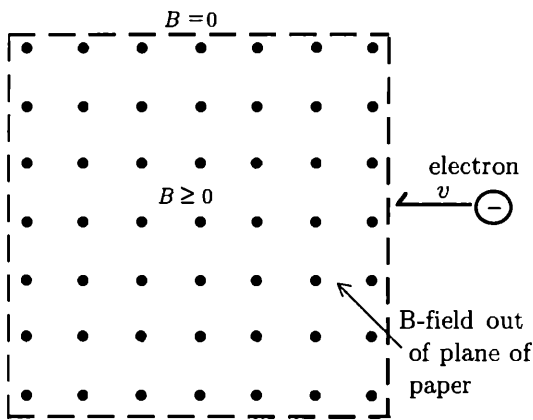
(1)

A possible trajectory of the electron with a velocity v very much *smaller* than that supposed in your initial trajectory (all other parameters held fixed). Label this trajectory v_{small} and explain your reasoning.

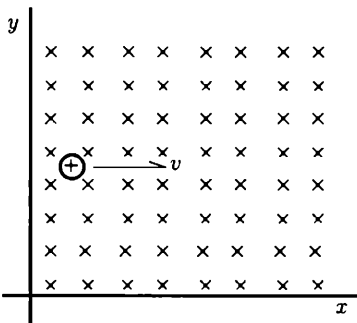
(2)

A possible trajectory that might be observed with a very much weaker B-field (all other parameters unchanged). Label this third trajectory B_{weak} . Explain your reasoning.
- (d)

Under the circumstances sketched, is it possible for the particle to make a complete circle within the B-field? Why or why not?



8.6 A positively charged particle moves in the positive x direction in a uniform magnetic field directed into the plane of the paper as sketched in the diagram. The resultant force on the particle can be made zero by introducing a uniform electric field of appropriate strength in the

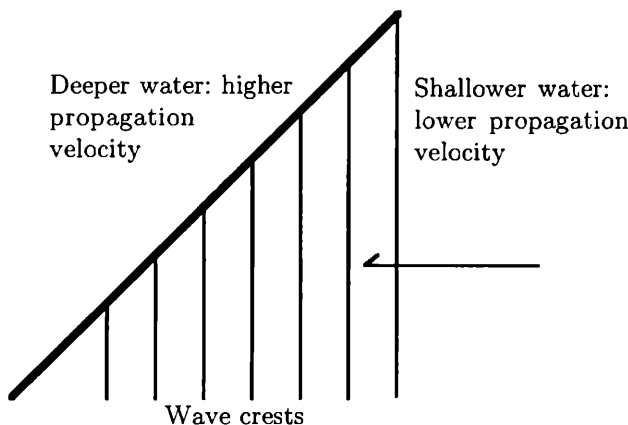


- (a) $+y$ direction.
- (b) $-y$ direction.
- (c) $+x$ direction.
- (d) $-x$ direction.
- (e) $+z$ direction (out of plane of paper).
- (f) $-z$ direction (into plane of paper).
- (g) None of the above; an electric field cannot cancel a magnetic field.

Chapter 9

Wave Phenomena

9.1 A straight wave train in a ripple tank is incident at an interface between shallower and deeper water.

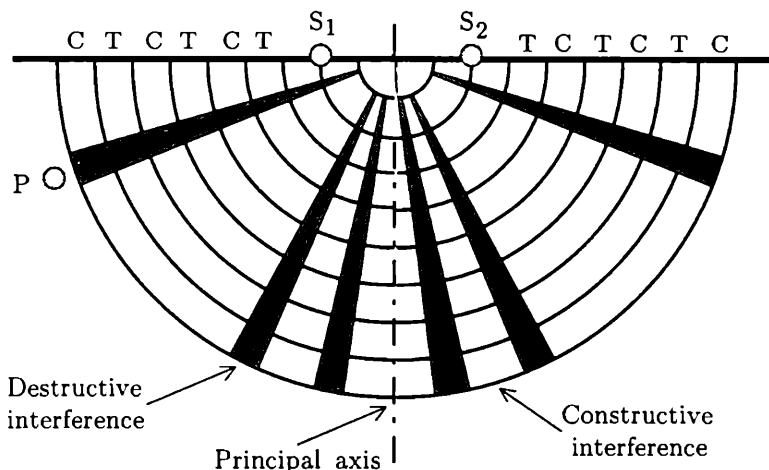


- (a) Add to the diagram (1) an incident ray, (2) a reflected ray, and (3) a refracted ray, labeling each one with the corresponding number.
- (b) Redraw the diagram and add (4) two or three reflected wave fronts and (5) two or three refracted wave fronts, labeling each set with the corresponding number.
- (c) If the frequency of the incident wave train is ν_i , how will the frequency ν_r of the reflected wave train compare with ν_i ? Will it be greater than, less than, or equal to ν_i ? How will the wavelength λ_r of the reflected wave compare with the wavelength λ_i of the incident wave? Explain your reasoning.
- (d) How will the frequency ν_t of the transmitted (or refracted) wave train compare with ν_i ? How will the wavelength λ_t of the transmitted wave compare with λ_i ? Explain your reasoning.

9.2 Is the following statement true or false? Explain your reasoning. "In a two-source interference pattern in a ripple tank, the nodes and antinodes all move out to

larger angles from the principal axis as the *frequency* of oscillation of the sources is decreased without change in separation of the sources."

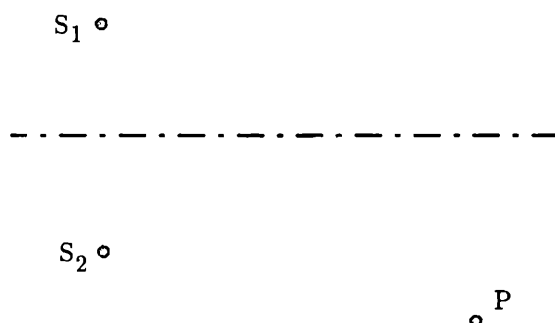
9.3 In this rough sketch of a two-source interference pattern in a ripple tank, the two sources S_1 and S_2 are oscillating in phase with each other. Crests and troughs are marked with the letters C and T respectively; the regions of constructive and destructive interference are indicated.



- Is the sketch internally consistent or inconsistent? Explain your reasoning. (Hint: Look at the comparative magnitudes of the spacing S_1S_2 between the sources and the wavelength of the ripples. Does the diagram make qualitative sense for the magnitudes indicated? Why or why not?)
- Consider the position denoted by the dot at P. What is the difference, at this point of observation, between the lengths PS_2 and PS_1 (i.e., the length $PS_2 - PS_1$) expressed in wavelengths λ of the ripples? Explain your reasoning.
- Sketch what this pattern might look like if the spacing S_1S_2 between the sources were decreased somewhat without any change in the wavelength of the ripples or the phasing of the sources.
- Sketch what this pattern might look like if the wavelength λ of the ripples were increased somewhat without change of spacing between the sources or in their phasing.
- Sketch what this pattern would look like if the source spacing and ripple wavelength were unchanged but the sources were oscillating exactly out of phase with each other.

9.4 In the diagram we are looking down on a two-source interference pattern in a ripple tank; the two point sources oscillate in unison. (The source positions S_1 and S_2 , the principal axis, and a position P are shown. The ripples themselves are not shown.) It is given that point P is located on the second line of destructive interference (antinode) away from the principal axis.

Use a ruler to measure the relevant distances directly on the diagram, and, from these distances, calculate the wavelength of the ripples making the pattern. Explain your reasoning.



9.5 Suppose that the diagram in Question 9.4 represents the essential features of a photograph of a two-source interference pattern in a ripple tank, a photograph that is *reduced* in scale relative to the tank pattern itself. The two sources S_1 and S_2 are known to be oscillating exactly *out* of phase with each other. The actual wavelength in the tank is known to be 2.0 cm. The observation point P is known to be on the second locus of destructive interference away from the principal axis; i.e., there is one other locus of destructive interference between this one and the principal axis.

- Using a ruler on the “photograph” and from information given above, calculate the *scale* of the photograph, i.e., how many centimeters in the photograph correspond to 1 cm in the ripple tank? Explain your reasoning.
- Calculate the actual separation between the sources in the tank. Explain your reasoning.

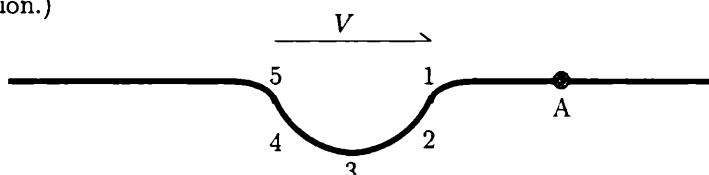
9.6 In describing sinusoidal wave trains, we always encounter expressions such as $\sin(2\pi x/\lambda)$ or $\cos(2\pi x/\lambda)$, where x denotes position along a coordinate axis and has dimensions of length. Explain in your own words where the $2\pi/\lambda$ comes from. Why not simply $\sin x$ or $\cos x$?

9.7 A horizontal, stretched string carries a sinusoidal wave train described by the equation $y = A \sin(ax + bt)$.

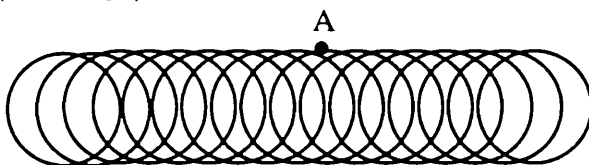
- In what direction along the x -axis is the wave moving? Explain your reasoning.
- What is the physical meaning of A ? Explain how you arrive at this interpretation.
- In terms of a and/or b , what is the wavelength of the wave train? Explain your reasoning.
- In terms of a and/or b , what is the frequency of the wave train? Explain your reasoning.
- In terms of a and/or b , what is the time taken for a particle of the string to go through one complete cycle of its up-and-down motion? Explain your reasoning.

- (f) In terms of a and/or b , what is the propagation velocity of the wave train? Explain your reasoning.
- (g) In terms of a and/or b , at what values of x are the deflections of the string zero at the instant $t = 0$? Show your mathematical reasoning.
- (h) What would be the effect of changing the plus sign in $(ax + bt)$ to a minus sign? Explain how you arrive at your answer.

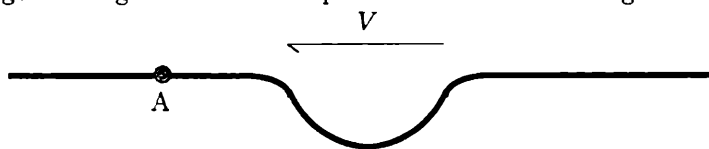
9.8 A pulse of the shape shown propagates to the right along a string with velocity V . Describe in detail the motion of particle A on the string as the pulse goes by. Be sure to indicate (1) the direction in which A is moving at various points along the pulse and (2) whether the velocity of A is increasing, decreasing, or zero at these various points. (Points 1 through 5 on the diagram are suggested as convenient references for the discussion.)



9.9 A pure rarefaction pulse (no compression phase) travels from right to left along this coil spring (or “slinky”).

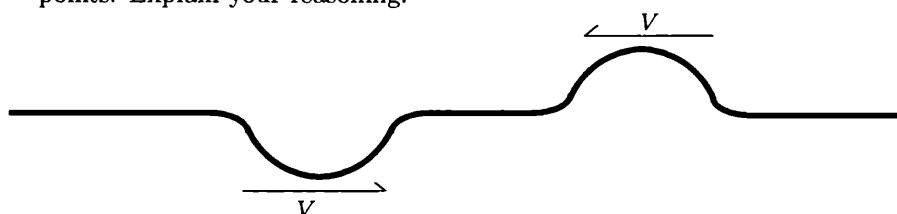


- (a) Describe what is happening to the coils as the pulse goes by, and describe the motion of point A on one of the coils, starting just as the leading edge of the pulse arrives at A until the instant the trailing edge passes. Include in your description the direction of motion of A, any changes in direction of motion, and the direction, as well as any changes in magnitude, of the velocity of point A. Does point A return to its initial position after passage of the pulse or is it permanently displaced? If A is permanently displaced, what altered shape of pulse would cause it to return to its initial position? Explain your reasoning as you go along.
- (b) What differences do you see between the particle motion you described in part (a) and the particle motion undergone by point A on the string in the case of passage of a negative *transverse* pulse such as the following?



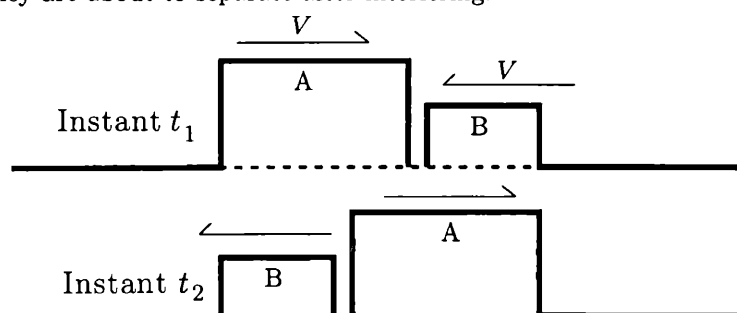
9.10 Consider the case in which two transverse pulses of identical shape but opposite phase travel in opposite directions along a string and interfere destructively.

- (a) At the instant the two pulses are exactly superposed, the deflection of the string is everywhere zero. Remember that each pulse is propagating energy and momentum in the direction in which it is traveling. What has happened to the momentum and energy at the instant of superposition? Note that their disappearance would be a violation of the conservation laws. In order to address this question, think carefully about the particle velocity at each point along the string at the instant of zero deflection, and indicate its direction at various points. Explain your reasoning.



- (b) Is the following statement true or false? “At the instant the two pulses cancel each other’s deflections destructively, all the energy of the two waves is present in the form of kinetic energy of particle motion of the string.” If the statement is false, alter it into a true statement; if you consider it true, explain why.
- (c) Perform an analysis parallel to that in (a) and (b) for the instant at which two identical positive pulses, propagating in opposite directions, superpose exactly and interfere constructively. In what form, at this instant, does one find the energy that is being propagated? Explain your reasoning.

9.11 For the sake of simplifying a useful exercise, let us imagine the following sharply stepped pulses A and B, propagating in opposite directions along a stretched string. (It is, of course, not physically possible to generate such idealized, perfectly rectangular pulses.) We imagine two photographs of the pulses: One taken at instant t_1 , just as the pulses are about to overlap (interfere), and the second taken at instant t_2 , just as they are about to separate after interfering.

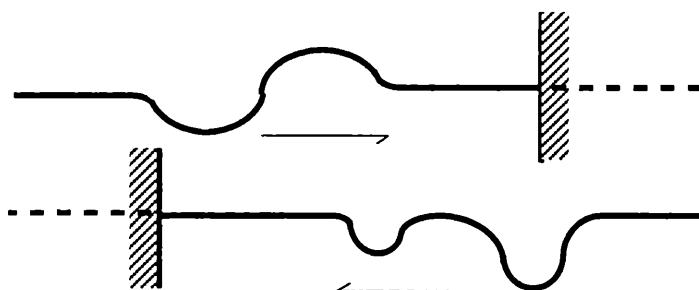


Take three instants of time, approximately equally spaced between instants t_1 and t_2 , and sketch the shape that the string would assume at each of those instants as pulses A and B overlap.

9.12 Transverse pulses of the shapes shown here propagate along strings in the directions shown. The end of each string is fastened to a rigid wall. Sketch the shape of the reflected pulse that would “emerge” from the imaginary (dashed) continuation

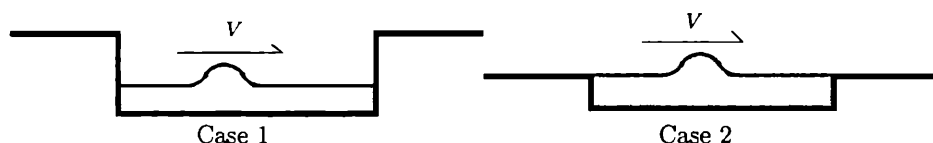
of the string on the other side of the wall in each case. Explain the reasoning behind your sketches.

Now suppose the ends of the strings were perfectly free instead of rigidly fixed. Sketch the pulses that would emerge from the other side of the free ends.



Note to the instructor: Problems like 9.12 should be posed with respect to reflections of transverse waves of other shapes at fixed and free boundaries and at boundaries between strings of different mass per unit length. Similar problems should be posed with respect to reflections of longitudinal wave pulses. The concepts do not become fixed without repeated opportunities to do the reasoning.

9.13 In a ripple tank, it is easy to make a straight wave pulse by quickly moving a rod either backward or forward near the water surface. Consider a ripple tank in which such a low amplitude pulse, in the form of a crest, is propagating from left to right and undergoes normal incidence at the end of the tank. (In these figures we are looking, from the side, at a cross section through the tank. Note that the tank has a fairly wide rim.) In case 1, the tank is only partly filled, and the ripple is incident at a vertical wall. In case 2, the tank is completely filled, and the ripple is incident at the rim.



- (a) In the light of what you have learned about reflection of waves at various boundaries, predict the shapes of the reflections that would be observed in the two cases illustrated. Would the reflections be identical or would they be different? Explain your reasoning in making your predictions.
- (b) Noting that the ripple is a transverse wave, compare the reflections predicted in the two cases above with (1) the reflections, at fixed and free boundaries, of transverse pulses on a string and (2) the reflections of longitudinal pulses, at fixed and free boundaries on a coil spring (or “slinky”). To which of the phenomena, (1) or (2), are the ripple reflections most nearly analogous?

9.14 Given a stretched string, describe how you would generate (1) a transverse pulse with only positive deflection (i.e., only a positive phase); (2) a transverse pulse

with only negative deflection (i.e., only a negative phase); (3) a transverse pulse with both positive and negative deflections (i.e., with both positive and negative phases). In your description, indicate what motion you would actually execute with your hand at one end of the string in order to produce the given pulse.

9.15 Given a stretched coil spring or “slinky” (say, lying on the table), describe how you might generate (1) a longitudinal pulse with only a positive phase (compression); (2) a longitudinal pulse with only a negative phase (rarefaction); and (3) a longitudinal pulse with both positive and negative phases. In your description, indicate what motion and displacement you would actually execute with your hand at one end of the slinky in order to produce the given pulse.

- (a) Now give a parallel description regarding generation of transverse pulses on a string: What motion and displacement do you execute with your hand at one end of the string to generate either a purely positive or a purely negative pulse?
- (b) Identify the very significant differences between the displacements you execute for transverse waves on the one hand and longitudinal waves on the other when it comes to generating pulses having only one phase, positive or negative.

9.16 Using appropriate sketches, describe what is meant by a “continuous wave train” in contrast to a single pulse. Describe what is meant by a “periodic” wave train. Describe what is meant by a “sinusoidal” wave train. What is the difference in meaning between the terms “periodic” and “sinusoidal”?

9.17 With the help of appropriate diagrams, define the terms “frequency, f ” and “wavelength, λ ” of a periodic wave train. Then, reasoning arithmetically from the definitions, establish the relationship among the three quantities V (velocity of propagation), f , and λ . (In other words, be able to reason out the relationship whenever you need it rather than memorizing it as a rigid formula.)

9.18 Define the concepts “ray” and “wave front,” using diagrams as well as words, and illustrating the concepts in the cases of both straight and circular pulses.

9.19 Using appropriate diagrams, define the concepts “angle of incidence” and “angle of reflection” for both straight and circular pulses incident at a straight rigid barrier. Define what is meant by “normal” and “glancing” incidence. Sketch and label the angles in wave front as well as ray representations. Sketch, in both ray and wave front representations, what happens as the angle of incidence is increased from normal to glancing.

9.20 Given a situation in which a straight wave pulse propagating in a region of deeper water (higher propagation velocity) is incident at a straight interface with a region of shallower water (lower propagation velocity). Sketch separate ray and wave front diagrams showing the incident, transmitted, and reflected pulses. Sketch such diagrams for the case in which the situation is reversed and the incident wave pulse arrives in the shallower region. In connection with your diagrams, define the concept “angle of refraction” (or “angle of transmission”).

9.21 In each of the instances and diagrams arising in the preceding question, sketch how the angle of refraction changes as the angle of incidence is varied from normal to glancing.

9.22 On the basis of your observations with circular wave pulses, sketch what happens to the wave front in reflections from a straight rigid barrier, i.e., show how the reflected wave changes as you move the center of the incident wave closer to, or farther from, the barrier.

9.23 Sketch what happens to wave fronts when a circular pulse is incident at a straight *refracting* interface, making diagrams that show different distances of the center of the incident circular pulse from the interface.

9.24 Suppose a straight sinusoidal wave train arrives at normal incidence to a straight barrier that is shorter than the length of the wave fronts (i.e., the unimpeded portion of the wave front can propagate past the barrier while part of the wave front is blocked). Sketch the pattern to be observed in the region beyond the barrier for the cases in which the wavelength λ of the wave train is (1) very short relative to the length of the barrier, (2) very long relative to the length of the barrier, and (3) of intermediate length.

9.25 A straight sinusoidal wave train arrives at normal incidence to a straight barrier that contains an opening of width D , and the waves are blocked except for passage through the opening. Sketch the pattern of wave fronts transmitted through the opening for different ranges of the ratio of wavelength to opening width λ/D (i.e., for small, large and intermediate values of λ/D).

9.26 With the help of appropriate sketches, define the concepts "refraction," "diffraction," and "interference" and explain how they differ.

9.27 You have probably seen a bow wave on the water surface generated by a rapidly moving boat or ship. Let us do some qualitative thinking about the circumstances in which bow waves are formed.

- (a) When a boat moves through the water, waves are invariably generated on the water surface. Suppose the boat moves very, very slowly. How does the velocity of the waves compare with the velocity of the boat under these circumstances? How do the waves behave relative to the boat? Does a bow wave form?
- (b) Suppose the boat now accelerates, slowly increasing its speed. What happens to the spacing between the leading waves and the bow of the boat? What speed must the boat attain so that the waves "stick" to the bow? What, in general, are the circumstances under which bow waves are formed? How does the shape of the waves formed by the slowly moving boat compare with the shape of the bow wave?
- (c) If the opportunity arises, reproduce the situation we have just been discussing by moving a vertically oriented pencil or rod through a ripple tank in the laboratory or a basin of water at home, and observe what happens at slow and at increasingly rapid motions. What is the shape of the ripples formed when the rod moves slowly? What is the shape of the bow wave? How do you account for the shape of the bow wave?
- (d) In the light of the preceding thinking and reasoning, try to visualize the effect of moving objects on the air that surrounds them. How would the compression

in front of a moving object behave relative to the object? Under what circumstances might a “bow wave” form? How would the effect of a baseball compare with that of a bullet?

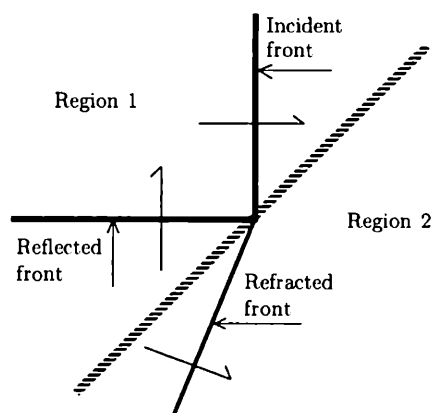
9.28 The shaded line represents a boundary between a deeper and a shallower layer of water in a ripple tank. (We are looking down on the surface of the water.) The velocity of propagation of ripples is higher in the deeper water. Suppose you were to take a pencil and move its point rapidly (faster than the propagation velocity of ripples on either side) in the water right along the boundary between the two depths, starting at the left-hand side.



Sketch the bow wave pattern that would be observed in each region (on either side of the depth change) at the instant the pencil point is at the right-hand end of the diagram. Explain your reasoning.

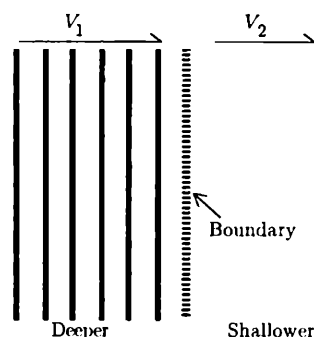
9.29 The diagram represents the arrival of a straight wave front at an interface between two regions of differing propagation velocity. The diagram also shows the refracted wave front.

- (a) Which region, 1 or 2, has the higher wave propagation velocity? Explain your reasoning.
- (b) Sketch a ray diagram that represents the same situation.

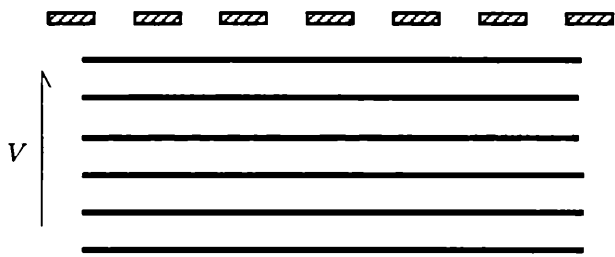


9.30 A straight wave train is propagating at velocity V_1 from left to right, in a ripple tank. The parallel solid lines represent crests of the train of ripples. The train is normally incident at a step in depth, the depth being shallower to the right and the propagation velocity V_2 lower. Let us suppose that V_2 is about half of V_1 (such a ratio would actually be hard to attain physically).

- (a) Sketch the spacing between the crests that would be observed on the right-hand side of the step change in depth and propagation velocity. Explain your reasoning.
- (b) Explain why it is that the *wavelength* of the wave train, rather than the frequency, changes under such circumstances. (Keep this fact in mind for future reference in the more subtle case of change in velocity of propagation of light.)



9.31 A straight wave train is propagating at velocity V in a ripple tank. The parallel solid lines represent crests of the train of ripples. The wave train is normally incident on a barrier that contains regularly spaced openings. The width of these openings is somewhat less than the wavelength of the wave train. The spacing between the centers of the openings is about twice the wavelength of the wave train. Diffracted wavelets, transmitted through the openings, superpose on the exit side of the barrier while the rest of the wave train is blocked.



Sketch the pattern of overlapping circular wave fronts you would expect to see on the exit side of the barrier under these circumstances as far as the first-order interference. Note that the pattern is very different from that of the two-source interference pattern. The superposition of circular wavelets leads to straight wave fronts propagating into the exit region. You can actually see the pattern of the straight wave fronts if you set the situation up carefully in a ripple tank. Make the wavelength as long as feasible and design the barrier accordingly. When you watch the pattern, let your eyes sweep along with the waves. You will then see the interference pattern that arises.

9.32 Shock cord is a very stretchable rubber “rope” that is used for highly flexible fastenings (a given segment can be stretched to more than twice its relaxed length). Suppose you start with a length of shock cord fastened at one end and pull on the other end, stretching the cord slightly under relatively low tension. You then make a transverse wave pulse by deflecting the end you are holding and note its propagation velocity V_1 . You then stretch the cord by quite a large amount, increasing its length (and the tension) significantly. You now observe a larger propagation velocity V_2 .

Explain, in terms of what you have learned about the velocity of transverse waves on a string, why V_2 is greater than V_1 . Make your explanation physical by considering the application of Newton’s second law to a small chunk of the string displaced from its equilibrium position, and not by simply referring to symbols in a formula for velocity of propagation. Note that two effects are involved (mass of the chunk and force applied), not just one. Be sure to account for each effect in your explanation.

9.33 A flexible coil spring (“slinky”) hangs from the ceiling. You produce a *transverse* pulse by deflecting the lower end. As the pulse propagates up to the ceiling, its propagation velocity changes continuously.

Predict, in terms of what you have learned about the velocity of transverse waves, how the propagation velocity changes on the way up. Does it increase or decrease? Explain why the velocity changes rather remaining the same, as it does in the case of a horizontal string. (Note that two effects are involved, not just one. Be sure to account for each effect in your explanation.)

9.34 The following questions extend to phenomena related to sound what you have learned about longitudinal waves by observing their behavior on a coil spring.

- (a) In the light of what you have seen of the generation and behavior of compression and rarefaction pulses on a slinky, sketch what you visualize might be happening to the air in a tube when a piston or diaphragm moves rapidly back and forth at one end.
- (b) Is it possible to make a sound pulse having only a compression phase and no rarefaction phase by moving the piston inward and then returning it to its initial position? Why or why not? How does this situation compare with the generation of pulses on the slinky? (Review Question 9.15.)
- (c) What variables (ordinate and abscissa) might you use to make a graph representing a sound pulse? There are several possibilities.
- (d) How would you imagine the interference of sound waves to take place? Suppose you had two audible point sources of sound, analogous to the situation with two point sources in the ripple tank (e.g., two tuning forks sounding in unison). How would you go about finding regions of constructive and destructive interference using only your own ears as detecting devices? That is, what positions would you explore and what would you listen for? Use a diagram to illustrate your answer.
- (e) Predict how the compression and rarefaction phases of sound pulses would be reflected from a rigid wall and explain your reasoning.

9.35 “Intensity (I)” of a wave train is defined as the energy transported (in the direction of propagation of the wave train) through a unit area per unit time. Thus the SI units would be watts per square meter per second [$\text{W}/(\text{m}^2)(\text{s})$]. Textbooks show that intensity I is directly proportional to the *square* of the amplitude A of the train.

- (a) How does the surface area of a sphere vary with the radius R of the sphere? In the light of this functional relation and the constraint of conservation of energy, how must the intensity of a spherically spreading wave train (e.g., a sound wave in air or water) vary with R if the wave spreads without dissipating any of its energy? Explain your reasoning. Qualitatively, how would you expect the intensity to vary with R , in comparison with the variation in the nondissipative situation, if there were small, continuous dissipation taking place (as is, of course, actually the case)?
- (b) In the light of the fact that intensity is proportional to the square of the amplitude, how would you expect the *amplitude* A of the wave train to vary with R in the ideal (nondissipative) situation in part (a)? Explain your reasoning.
- (c) Consider a train of water waves on the surface of a pond, generated, for example, by moving a stick up and down at the source. In this case, the wave train is spreading *cylindrically* rather than spherically as in parts (a) and (b). How does the area of a cylindrical surface vary with the cylindrical radius R ? In the light of this functional relation and the constraint of conservation of energy, how must the intensity of the cylindrically spreading wave train vary with R in

the ideal (nondissipative) case? How must the amplitude A of the cylindrically spreading wave train vary with R ? Explain your reasoning.

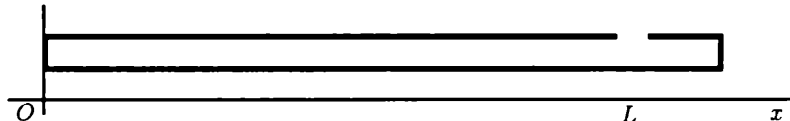
- (d) Making up some examples of your own with numerical ratios of radial distances, compare the decay of amplitude with distance in the spherical wave with the decay of amplitude with distance in the cylindrical wave, and illustrate the enormous quantitative difference between the effects of spreading in the two situations. What is the origin of the difference in the rate of decay with distance in the two different geometries? Explain your reasoning.

9.36 Given the following equation for the n th harmonic (n th eigenfunction) of a standing wave on a string with fixed ends. (Our clock starts at $t = 0$.)

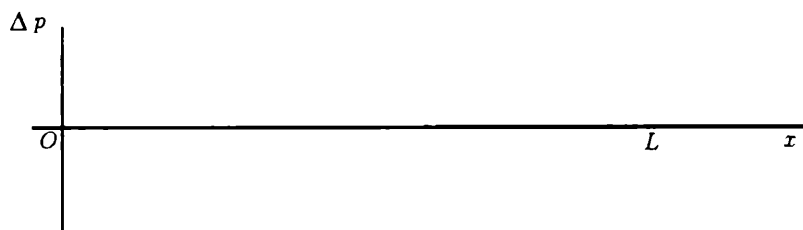
$$y_n = 2A \sin \frac{n\pi x}{\lambda_n} \cos \omega_n t$$

At what subsequent clock readings t will the string be completely flat (i.e., straight and undeformed)? Present your analysis in clear, sequential mathematical steps with explanations of reasoning as you go along. Express your final result in terms of the period T_n of the oscillation. (Note: If you encounter a new sequence of integer numbers, you will need a symbol other than n , since n has been preempted for the designation of eigenstates.)

9.37 An organ pipe has its closed end at the origin ($x = 0$) and its mouth at $x = L$ as sketched.



- (a) It is asserted that in a specific case of resonance in the pipe, the wave number k has the value 2.28 ft^{-1} and the angular frequency ω has the value 2510 s^{-1} . Is this combination of numerical values physically reasonable or unreasonable for a sound wave in air at room temperature? Explain your reasoning.
- (b) If the wave described in part (a) is the second harmonic of the pipe, what must be the length L of the pipe? Find the numerical value and explain your reasoning.



- (c) Using the departure from atmospheric pressure (Δp) as the dependent variable, sketch the standing wave form in (b) on the above set of coordinates (according to the usual convention for depicting standing waves). Indicate the locations of pressure nodes and antinodes.

- (d) Which of the following mathematical expressions would most conveniently describe the standing wave pressure variation in the pipe with respect to x and t as independent variables, given the coordinate system in part (c)? Circle your choice and explain the reasoning behind it. If more than one of the equations would be equally convenient, say so and explain why.

$$\Delta p = \Delta P_{\max} \cos \left[\frac{2\pi x}{\lambda} \pm \omega t \right]$$

$$\Delta p = \Delta P_{\max} \sin \frac{2\pi x}{\lambda} \sin \omega t$$

$$\Delta p = \Delta P_{\max} \cos \frac{2\pi x}{\lambda} \sin \omega t$$

$$\Delta p = \Delta P_{\max} \sin \frac{2\pi x}{\lambda} \cos \omega t$$

9.38 Following are some functions of the form $f(x \pm Vt)$. Examine each function and assess whether it might be useful for describing a physically possible propagating wave form. Explain your reasoning.

$$y = A(x - Vt)$$

$$y = A(x + Vt)^2$$

$$y = A\sqrt{x - Vt}$$

$$y = A \ln(x + Vt)$$

$$\left\{ \begin{array}{ll} y = 0 & \text{for } x < 0 \\ y = A \exp^{-(x-Vt)} & \text{for } x \geq 0 \end{array} \right\}$$

The last form is actually a good approximation for the shape of the shock wave emitted in an underwater explosion. In the real situation, the wave, of course, spreads out radially rather than being confined to one dimension. In the three-dimensional situation, the amplitude varies (to a first approximation) inversely as the radius. Write the equation appropriate to the spherically spreading wave.

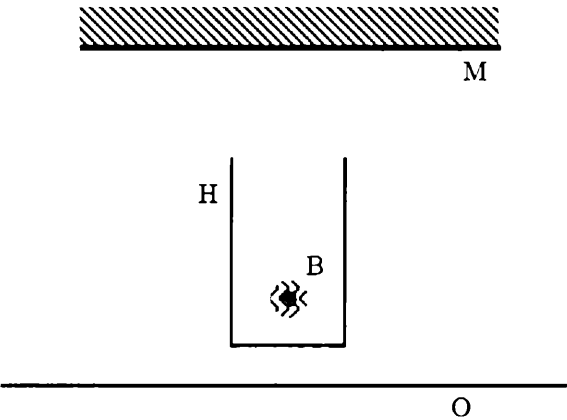
9.39 Describe *qualitatively* what actually happens to the energy being transported by a transverse wave on a string when the wave is incident at a *real* wall. Is *all* the energy reflected? Why or why not? What happens in the wall? The phenomena taking place in the wall are actually very complex, but try to visualize at least some of what might happen.

9.40 Can you generate, on a stretched string, a wave that would transport angular momentum as well as linear momentum? Why or why not? If you can generate such a wave, describe how you would do so and what you would see happening along the string.

Chapter 10

Images with Mirrors and Lenses

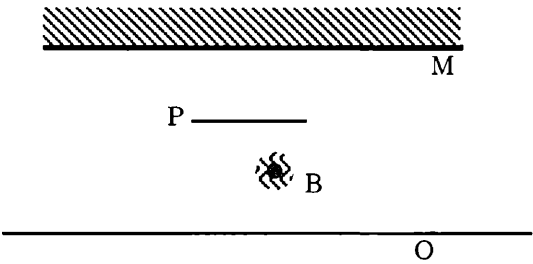
10.1 In the following diagram, M is a plane mirror; B is a very small, bright light bulb that can be treated as a point source of light; H is an opaque housing that does not transmit light; and O is a line anywhere along which an observer can stand to try to see the image of the light bulb in the mirror.



- (a) By using relevant rays of light, determine the locations along line O from which the image of B is visible in the mirror and the locations from which it is not visible. Mark these regions accordingly along line O, and explain the reasoning you used in drawing the rays.
- (b) Explore how moving B around within the housing H would affect the regions you have mapped out.

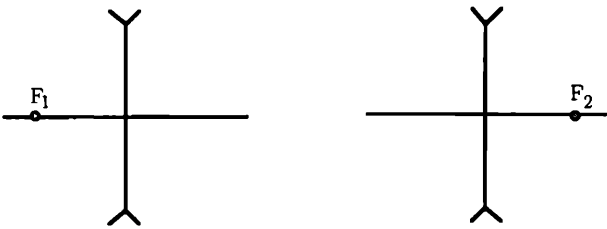
10.2 In the following diagram, M is a plane mirror; B is a very small, bright light bulb that can be treated as a point source of light; P is an opaque plate that does not transmit light; and O is a line anywhere along which an observer can stand to try to see the image of the light bulb in the mirror.

- (a) By using relevant rays of light, determine the locations along line O from which the image of B is visible in the mirror and the locations from which it is not visible. Mark these regions accordingly along line O, and explain your reasoning.
- (b) Explore how moving B around in the region behind plate P would affect the regions you have mapped out.

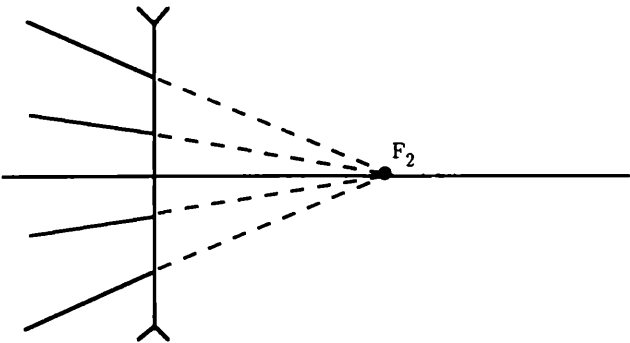


Note to the instructor: Many of the following problems on images with lenses would have counterparts with concave and convex mirrors, probing similar levels of understanding. The problems could simply be rephrased accordingly.

10.3 In the two diagrams, give definitions (separately) of each of the principal focal points F_1 and F_2 of a thin diverging lens by (a) drawing rays that define the points and (b) describing in your own words how the rays are drawn.



10.4 Point F_2 is a principal focus of the diverging lens, which intercepts a cone of rays converging toward point F_2 from the left.

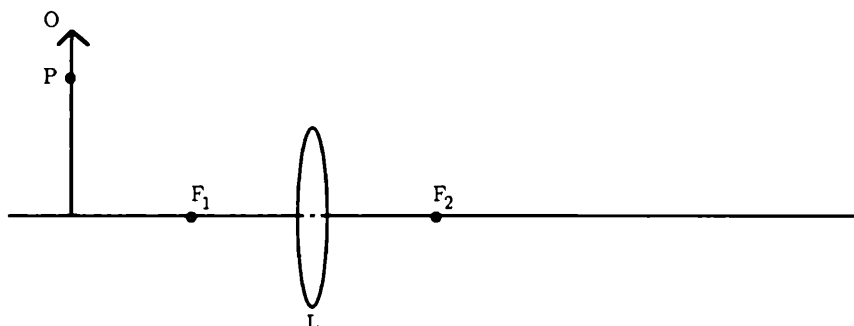


Add to the diagram and label the following additional lines:

- (a) A wave front associated with the incoming converging rays.

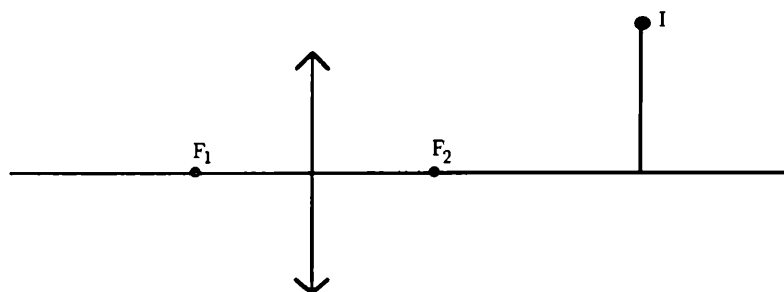
- (b) The rays that emerge from the lens on the right-hand side.
- (c) A wave front associated with the emerging rays.

10.5 A thin converging lens L has principal foci at F_1 and F_2 . The lens forms an image (not shown) of object O .

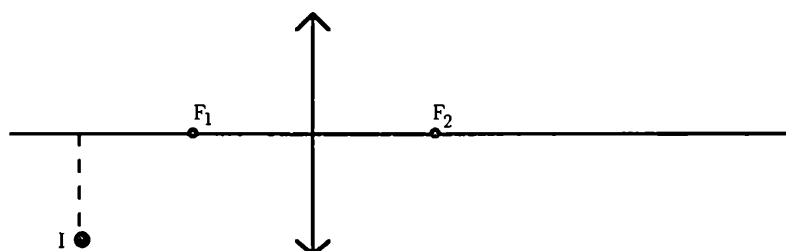


- (a) Is the image formed by the lens that of the *entire* object or does the image encompass only *part* of the object? Explain your reasoning.
- (b) Consider the point P . Like any other point on the object, this point sends out light rays in all directions. Determine what happens to light emanating from point P : What happens to the rays that are not intercepted by the lens? What happens to the rays that *are* intercepted by the lens? Give the answers by drawing appropriate rays on the diagram, and then describe the results in words.
- (c) Sketch a few wave fronts for rays from point P both before and after interception by the lens.

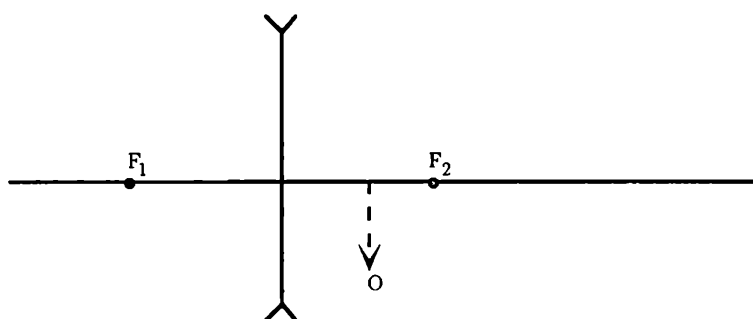
10.6 A thin converging lens with principal foci at F_1 and F_2 forms a real image I at the position shown in the diagram. Draw the three principal rays that will permit you to determine where the object must have been in this case, and describe in words how each ray is drawn.



10.7 A thin converging lens forms a virtual image I at the position shown in the following diagram. Draw the three principal rays that determine where the object must have been in this case, and describe in words how each ray is drawn.



10.8 The dashed arrow and symbol O mark the location of a real image that is being formed by a converging lens some distance off to the left. The light coming from the converging lens is intercepted by the diverging lens shown in the figure. The arrow O is therefore a *virtual* object for the diverging lens, the principal foci of which are located at the points labeled F_1 and F_2 , respectively.

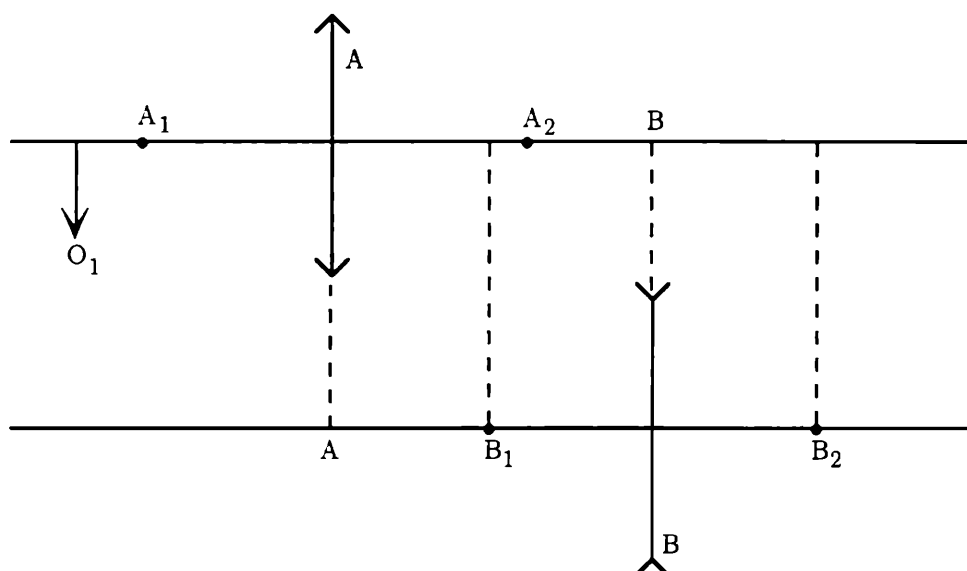


- Draw the three principal rays so as to establish the location and size of the final image.
- Describe in your own words how each of the three principal rays is drawn.
- Indicate whether the final image is real or virtual and explain how you arrived at your conclusion.
- Sketch a few wave fronts for the rays intercepted by the diverging lens and for the rays emerging from the diverging lens.

10.9 A thin converging lens is placed at position A on an optical bench as shown in the *upper* part of the following diagram. The lens has a focal length of 3.00 cm, and its principal foci are shown at the positions marked A_1 and A_2 , respectively. An object O_1 is placed at a position 4.00 cm to the left of the lens. (The diagram is drawn to scale.)

- On the upper part of the diagram, draw the principal rays that locate the image formed by the converging lens. Describe in your own words how each ray is drawn (put this writing elsewhere on the page; do not clutter the diagram).
- Use the relevant lens equations to calculate the position of the image and the lateral magnification.

- (c) Interpret the results you obtained in parts (a) and (b): Do your graphical and algebraic results agree or disagree? Is the image real or virtual?



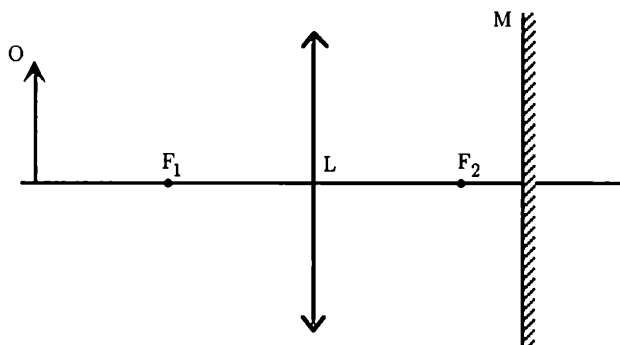
Without altering the situation established in part (a), a thin diverging lens is now placed at position B on the optical bench just 5.00 cm to the right of lens A, intercepting the cone of light coming from lens A. The focal length of lens B is 2.50 cm, and the principal foci are marked B_1 and B_2 . (The diverging lens is shown in the lower part of the diagram, but this is done to avoid cluttering the upper part. The two lenses are actually in line on the same optical bench.) The image formed by lens A now becomes the object for lens B.

- (d) Is the object for lens B real or virtual? Explain your answer.
- (e) Complete the diagram, drawing the three principal rays establishing the final image formed by lens B. Describe in your own words how each ray is drawn. (If you use a second color for the principal rays of the diverging lens, you may complete the diagram in the upper portion where you have already established the image formed by lens A. If you are using a single color, complete the drawing in the lower portion of the diagram.)
- (f) Use the lens equation to calculate the position of the final image and the lateral magnification produced by lens B.
- (g) Interpret the final results: Is the final image real or virtual? Do your graphical and algebraic results agree or disagree?

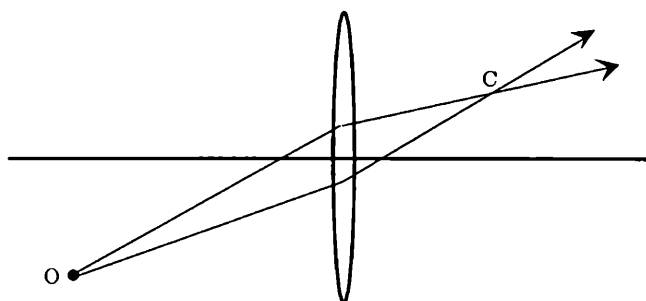
10.10 In the following system of object O and converging lens L, the plane mirror M reflects the light emerging from the lens.

- (a) Carefully draw a ray diagram establishing the final location of the image being formed by the system. Is the object for the mirror real or virtual? Explain your reasoning.

- (b) Is the final image real or virtual? If it is real, how do you reconcile this with the fact that the images you normally see in a plane mirror are always virtual?
- (c) Describe what you might do with either the position of object O or the position of the mirror to make the object for the mirror real instead of virtual. Where would the final image formed by the system then be located?



10.11 Point O is a source of light. Two rays from O are shown passing through the thin converging lens and crossing each other at point C.



- (a) Find the two principal foci of the lens by drawing appropriate rays, and label them F_1 and F_2 , respectively. Explain your reasoning.
- (b) Draw a ray from O to the extreme upper end of the lens and show the direction in which it emerges from the lens. Explain your reasoning.
- (c) Where would you place a screen so that the image of O would be in focus on it? Explain your reasoning.
- (d) Show on the diagram where you would place your eye, and the direction in which you would look, to see the image of O without the presence of a screen. Explain your reasoning.
- (e) Suppose the lens is replaced by another lens having the same focal length but a larger diameter. Will the location of the image along the principal axis change? Why or why not? If it changes, how will it change? Will the distance of the image from the principal axis change? Why or why not? If it changes, how will it change? Will the brightness of the image change? Why or why not? If it changes, how will it change?

10.12 Consider the following statement: “If it were possible to make *perfectly* smooth, *perfectly* spherical lenses, one would then have lenses that would not have to be corrected for aberrations.” Is this statement true or false? Indicate your line of reasoning.

10.13 Consider the following statement: “In using a slide projector in which a large image is projected on the screen, the slide must be located just *inside* the principal focus of the lens (i.e., slightly *closer* to the lens than the focal point).” Is this statement true or false? Explain your reasoning with the aid of an appropriate diagram.

10.14 Consider the following statement: “When a camera is used to take pictures, the film must be located at the principal focal plane of the lens for very distant subjects and somewhat *closer* to the lens than the principal focal plane for closer subjects.” Is this statement true or false? Explain your reasoning with the aid of an appropriate diagram.

10.15 Consider the following statement: “As we come closer to an object, the lens system of our eye must become *more strongly converging* to focus the image sharply on the retina.” Is this statement true or false? Explain your reasoning with the aid of an appropriate diagram.

10.16 Let us examine a very basic aspect of the everyday use of cameras and slide projectors.

- (a) Suppose you are using a camera and wish to have a larger image of a distant object than you are obtaining with the lens currently in use. Would you change to a lens with a longer or a shorter focal length? Explain your reasoning.
- (b) Suppose you are using a slide projector and wish to obtain a larger image on the screen. You cannot achieve this by moving the screen farther from the projector because you are already using the entire length of the room. Would you change to a lens with a longer or a shorter focal length than the one you are using? Explain your reasoning.

It is possible to address these two questions simply by drawing and interpreting appropriate geometrical diagrams. It is a valuable exercise, however, also to address the questions mathematically (by appealing to the lens equation and the expression for lateral magnification) and proving your contentions analytically. This is a nice exercise in elementary mathematical physics. Be sure to verify that your analytical solution agrees with your geometrical one. [Hint: It is important to realize and keep in mind that in part (a) the object distance is essentially fixed while in part (b) the image distance is essentially fixed.]

Note to the student: In the following multiple-choice questions, mark those statements that are correct. *Any number* of statements may be correct, *not* just one. You must consider each statement on its merits.

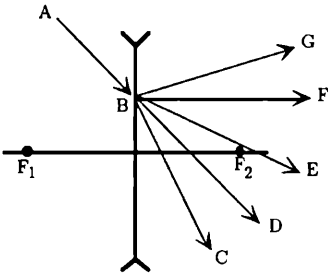
10.17 A simple glass lens exhibits what is called “chromatic aberration”: i.e., its focal length for red light is longer than its focal length for blue. This behavior of the lens can be explained on the basis that

- (a) the resolving power of a lens is greater for blue light than for red.

- (a) the resolving power of a lens is greater for blue light than for red.
- (b) longer wavelength rays are refracted more strongly at the edges of the lens than nearer the center.
- (c) shorter wavelength rays are refracted more strongly at the edges of the lens than nearer the center.
- (d) red light is retarded less strongly than blue on passing from air to glass.
- (e) red light is retarded more strongly than blue on passing from air to glass.
- (f) the diffraction introduced by the lens separates the colors.
- (g) the lens is not perfectly spherical.
- (h) the lens surface is not as smooth as is necessary for perfect focussing.
- (i) None of the above.

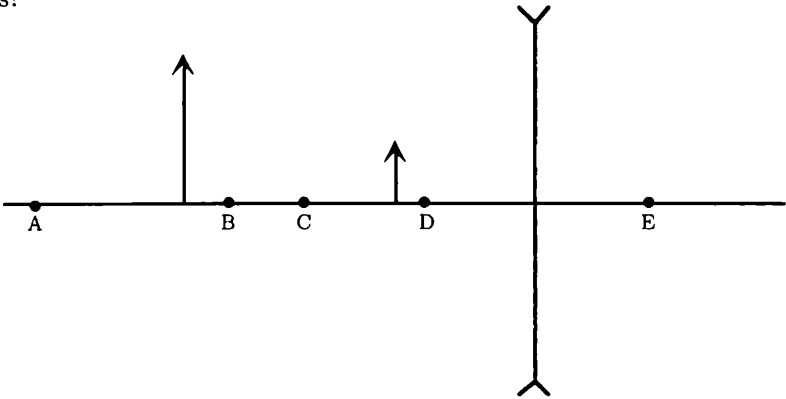
10.18 This diverging lens has principal foci at F_1 and F_2 . A ray of light AB is incident from the left at point B on the lens. The emerging ray is best represented by

- (a) ray BC.
- (b) ray BD.
- (c) ray BE.
- (d) ray BF.
- (e) ray BG.



Explain your reasoning.

10.19 The arrows in the following diagram represent object and image formed by a diverging lens. Which of the points marked A, B, C, D, E is closest to a focal point of the lens?

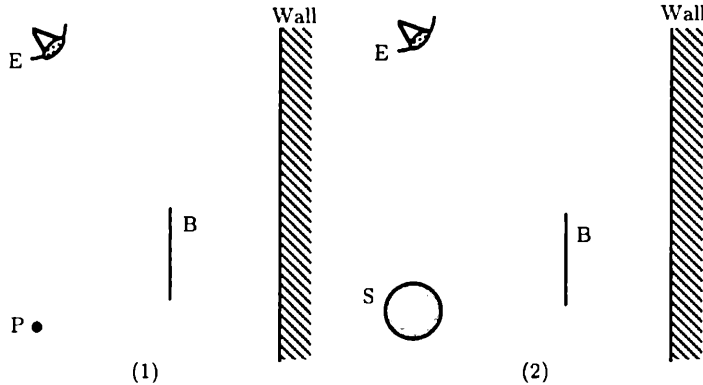


- (a) Point A.
 - (b) Point B.
 - (c) Point C.
 - (d) Point D.
 - (e) Point E.
 - (f) Both points D and E.
- Explain your reasoning.

Chapter 11

Geometrical and Physical Optics

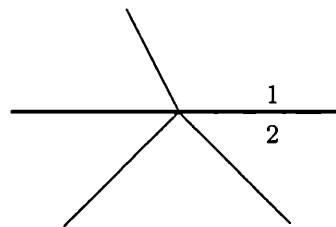
11.1 In diagrams (1) and (2), a bright source of light casts the shadow of an opaque obstacle B on a wall that is a diffuse reflector of light (not a mirror). The eye of an observer is shown at E. In (1) the source of light, P, is to be treated as a point source of negligible spatial extent, sending out light in all directions. In (2) the source of light, S, also sends out light in all directions, but the source has substantial size, as shown.



- (a) In each diagram, draw rays of light that originate at the source and reach the eye of the observer, setting the boundaries of the shadow that the observer sees on the wall. Interpret the diagrams so as to explain how it comes about that the shadow in (1) exhibits only one region, uniform in darkness, while the shadow in (2) consists of two regions, a dark inner region (called the “umbra”) and a lighter outer region (called the “penumbra”).
- (b) When the shadow of object B is cast by the sun, the shadow is quite sharp when B is relatively close to the wall and is much fuzzier, with obvious umbra and penumbra, when B is relatively far from the wall. How do you explain the difference between these two situations?

11.2 In an illuminated room, you can see any wall of the room from any location in which you stand. Is there any reflection involved in the interaction between light and the walls? If not, how do you account for being able to see the wall at all? If so, how do account for the fact that you do not see images of other objects in the room?

11.3 If we see a reflection of the moon or the sun in a very still lake, we see a relatively sharp outline with perhaps a few wiggles around its periphery. If, however, the water surface is ruffled by the wind, we see a broad streak stretching toward us along the surface. Explain these observations in terms of what you now know about the reflection of light. Use sketches to assist and clarify your explanation.



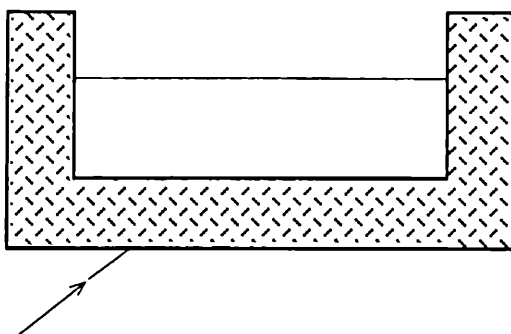
11.4 A light ray is incident at an interface between two media (1 and 2) and is partially reflected and partially transmitted as implied in the diagram.

- Identify the incident, reflected, and transmitted rays and explain how you made the identification. Draw arrow heads on each ray indicating the direction of propagation of the light.
- Label the angles of incidence, reflection, and transmission (refraction).
- In which medium does light have the higher velocity of propagation? Explain your reasoning.
- Which of the two media has the higher index of refraction? Explain your reasoning.
- Draw short lines on each ray showing the orientation of the wave fronts associated with each ray.
- Does the frequency of the light change on passing from one medium to the other? If it changes, explain, in terms of what happens at the interface, why it changes, and indicate whether it increases or decreases. If it does not change, explain why it doesn't.
- Does the wavelength of the light change on passing from one medium to the other? If it changes, explain, in terms of what happens at the interface, why it changes, and indicate whether it increases or decreases. If it does not change, explain why it doesn't.

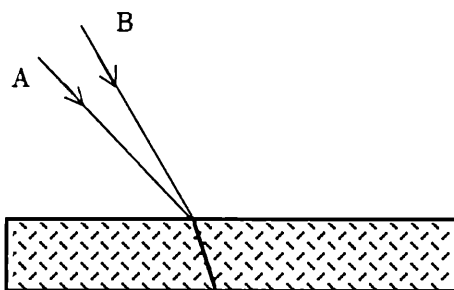
11.5 A layer of water is contained in a glass tank as shown. (The index of refraction of glass is greater than that of water.) A ray of light is incident at the bottom of the tank as shown.

- Sketch the ray transmitted into the glass and then into the water and out of the water into the air.

- (b) Now assume that the water is removed from the tank. Sketch a ray incident, on the bottom of the tank as before, transmitted into the glass, and then into air instead of water. What differences are there between the rays in parts (a) and (b)? Explain.

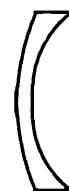


11.6 Two very narrow rays of light of different colors are incident on a glass plate at different angles. (The difference in angle is exaggerated for the sake of clarity.) It so happens that the two rays coalesce into one ray on entry into the glass, as shown in the diagram.



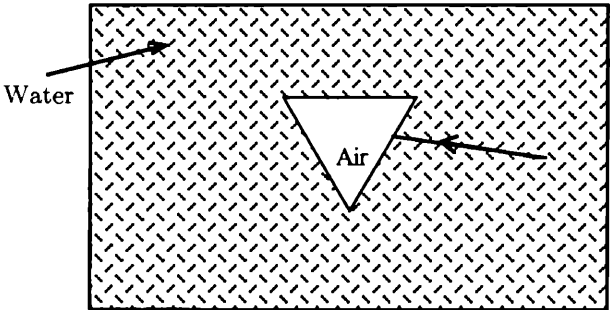
- Which of the two rays has the higher index of refraction in glass? Explain your reasoning.
- Sketch the rays reflected from the bottom surface of the glass and emerging back into the air through the upper surface of the glass. Explain your reasoning.
- Suppose the angle of incidence of rays A and B were increased without changing the angle between them. At what angle of incidence at the upper surface of the glass would a ray of either A or B, reflected from the bottom surface, undergo total internal reflection within the glass? What would simultaneously be happening to the ray of the other color? Explain your reasoning.

11.7 A lens of glass or transparent plastic is molded into the shape shown in the figure. A parallel beam of light, parallel to the principal axis, is incident on this lens from the left. Will the beam emerging from the lens be converged or diverged, or will it remain parallel to the principal axis? Explain your reasoning.

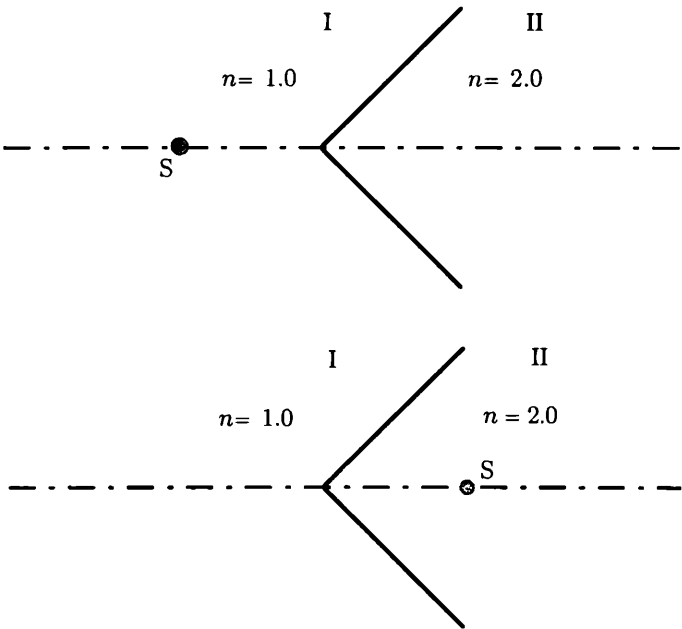


11.8 A hollow “air prism” is made with walls of thin transparent plastic and is sealed to be watertight. The prism is immersed in water as shown. A ray of light enters the prism from the water on the right.

Sketch a possible refracted ray entering the prism, passing through it, and emerging back into the water through another wall. Show normals to the surface and reflected rays as well the refracted ray at each interface. (The walls of the prism are to be treated as having negligible thickness and playing no significant role in the path of refraction.)

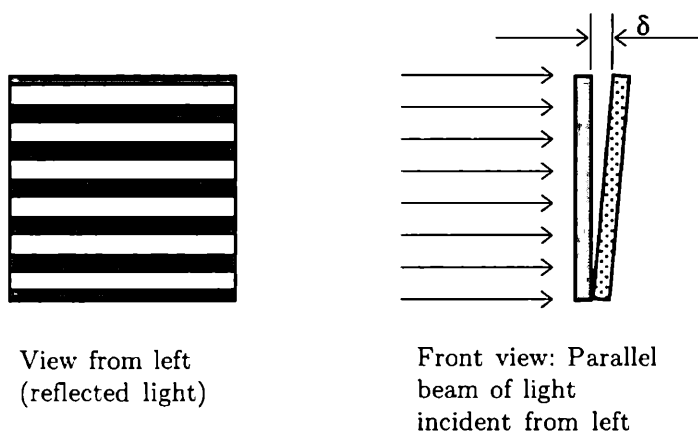


11.9 In the following diagrams, region I consists of air with an index of refraction of 1.0, and region II consists of a solid transparent material with index 2.0. The sides of the transparent solid are perpendicular to each other at the vertex, and the dashed principal axis bisects this right angle. In the first diagram, a point source S, sending out light in all directions, is located on the axis in region I, and, in the second diagram, it is located on the axis in region II. (In the following questions, confine your investigations to the plane represented in the diagrams, i.e., the plane of the paper.)



- (a) In the upper diagram, is there a region within region II in which a detector would “see” two apparent point sources of light? If you conclude there is no such region, explain your reasoning. If you conclude there *is* such a region, map out its boundaries in the plane of the diagram, explaining how you arrived at this conclusion and how you established the boundaries.
- (b) In the lower diagram, is there a region within region I in which an observer would see two apparent sources of light? If you conclude there is no such region, explain your reasoning. If you conclude there *is* such a region, map out its boundaries in the plane of the diagram, explaining how you arrive at this conclusion and how you establish the boundaries.

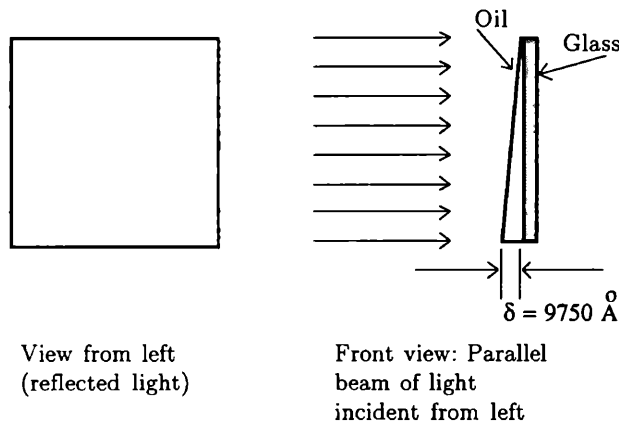
11.10 A thin wedge-shaped film of air is present between two glass plates as shown in the diagram. The thickness of the film is essentially zero where the plates are in contact at the bottom of the figure, and the thickness at the top has the unknown value denoted by δ . The plates are illuminated from the left by a parallel beam of monochromatic light with a wavelength of 5400 angstroms. When the reflected light is viewed from the left, one sees the interference pattern shown at the left.



- (a) What is the color of the light being used? (Cite your source of information.)
- (b) Given the nature of the observed pattern, what can you say about the flatness and smoothness (or the departure from flatness and smoothness) of the surfaces of the two plates? What might be the nature of the pattern if the surfaces were not very flat? Explain your reasoning.
- (c) Calculate the thickness δ of the air film at its upper end, explaining your reasoning and giving the result in both angstrom units and centimeters.

11.11 A thin film of oil clings to the left-hand side of a very flat and smooth glass plate, as shown. Under the influence of gravity, the oil film takes on a wedge-like shape since it tends to thicken at the bottom. The thickness of the film is essentially zero at the top and 9750 angstroms at the bottom. The indices of refraction of the oil and the glass are 1.4 and 1.6, respectively.

A parallel beam of red light (wavelength 6500 angstroms) is incident from the left. The film is viewed from the left, and the reflected light forms an interference pattern. Determine the location and the number of bright and dark bands that will be observed in the interference pattern and sketch the predicted pattern in the blank square at the left-hand side of the figure. Be sure to explain your reasoning and mark which bands are bright and which are dark.



11.12 Consider the following statement: “As an optical grating is made finer (i.e., as spacing between lines is decreased), the various orders of spectra (1st, 2nd, 3rd, etc.) all lie closer to the principal axis.” Is this statement true or false? Explain your reasoning.

11.13 Consider the following statement: “The fact that a lens is capable of producing a real image by refracting light and focussing it on a screen demonstrates that light must be a wavelike phenomenon.” Is this statement true or false? Explain your reasoning.

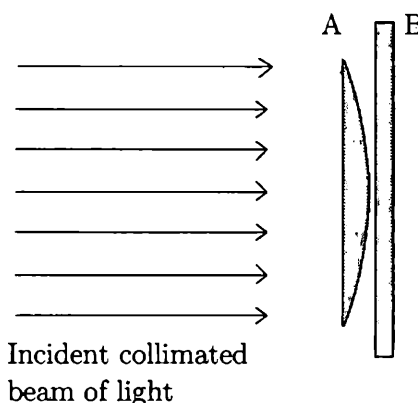
11.14 A narrow beam of white light is spread out into a spectrum on passage through a grating and on passage through a prism.

- Sketch the spectrum that would be observed in the case of the grating with emphasis on the location of the red and blue ends of the spectrum with respect to the principal axis of the system.
- Make a similar sketch for the case of the spectrum produced by the prism.
- State clearly and concisely in your own words what *inferences* are to be drawn from these observations about comparative properties of red and blue light. That is, what does each experiment tell us about properties such as wavelength and/or velocity of propagation of the different colors? What does each experiment *not* tell us about wavelength and/or velocity of propagation of the different colors?

11.15 Suppose a glass (index of refraction 1.52) converging lens with a given focal length is placed in water (index of refraction 1.33). What do you predict will happen

to the focal length of the lens while it is located in the water? Will it increase, decrease, or remain unchanged? Explain your reasoning.

11.16 In the so-called “Newton’s rings” experiment, a collimated (parallel) beam of monochromatic light is incident on the system illustrated: A plano-convex glass lens (A) rests against a glass plate (B) with air between the two pieces of glass. A and B are made of the same kind of glass. Both the light reflected from this system and the light transmitted through it form patterns consisting of concentric bright and dark rings. We shall concentrate on the reflection pattern. This pattern begins with a *dark* circular region at the center, followed by a next bright ring, etc. (If you have not seen a Newton’s rings demonstration, you should ask to see one.)



Early in the 19th century, a controversy raged over the comparative merits of the wave and particle models of light. Opponents of the wave model pointed to the dark central region of the Newton’s rings reflection pattern as a serious failure of the model. Their reasoning was that since the air film at the center, where the two glasses are in “contact,” must be extremely thin relative to the supposed wavelength of the light, there should be zero phase shift between the light reflected from the lens-air interface and that reflected from the air-plate interface, since the back and forth travel time in the film would be essentially zero. If interference is taking place between the two reflected beams, it should be constructive rather than destructive, and the central region of the pattern should be bright rather than dark.

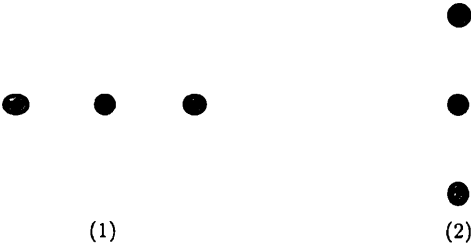
It occurred to Thomas Young, vigorous proponent of the wave model that, by analogy to the behavior of waves in mechanical systems, light waves might undergo either a zero or a complete 180° phase shift on reflection at interfaces, depending on whether the index of refraction increases or decreases at the interface. (Young conceived of change of index as analogous to relatively fixed or relatively free boundaries for waves on strings or springs.)

Young proceeded to alter the setup illustrated above by making lens A out of crown glass with an index of refraction of about 1.5 and the plate B out of flint glass with an index of about 1.7. Instead of air between the two pieces of glass, he put a film of sassafras oil, having an index of about 1.6. The pattern on reflection now showed a bright, instead of a dark, central region. The experiment proved to be a compelling victory for the wave model.

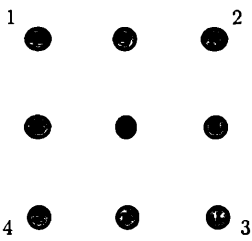
- (a) Interpret Young’s clever experiment in your own words. What was the point of using the sassafras oil? How did the experiment support Young’s hypothesis concerning phase shifts? What information does it *not* give concerning the phase shifts? How does the overall combination of observations, including the earlier ones, support the idea that the observed patterns are interference patterns and thus support the wave model?

- (b) Does Young’s experiment give any direct information about “what is waving” in a “light wave”? Does it give any information as to whether the wave is longitudinal or transverse? Explain your reasoning.
- (c) In the foregoing sequence of experiment and reasoning, identify the observations and the inferences drawn from the observations, distinguishing clearly between the two modes. What information that might be desired *cannot* be drawn from these specific observations?
- (d) Why do you think it was customary in the 1800s to refer to media with higher indices of refraction as optically “more dense” than media with lower indices?

11.17 When a narrow, circular beam (pencil) of monochromatic light is incident on a diffraction grating whose lines run *vertically*, the resulting diffraction pattern (to the first order) is that shown in figure (1) on the left. If the lines of the grating are oriented *horizontally*, one obtains the pattern shown in (2) on the right.



If one now crosses the two gratings one on top of the other (i.e., the lines of one run horizontally while those of the other run vertically), one obtains the following pattern, with a total of *nine* spots instead of the six that might be expected from (1) and (2).



- (a) How do you account for the sudden appearance of the four new spots (1, 2, 3, 4) on the diagonals at the corners of the square, spots that were not present at all with the individual gratings? [Hint: Draw a picture of the overlapping grating lines, and examine the pattern of openings carefully for lines of openings (other than horizontal and vertical) that would lead to diffraction patterns.]
- (b) Note the relative spacing between the array of openings that you identify as causing the new spots and show that these new spots must lie exactly on the corners of the square.

- (c) Suppose you rotated the crossed gratings rapidly around the central axis in the plane of the paper. What would be the resulting pattern of bright and dark areas?

Note to the instructor: The circular pattern resulting in part (c) of Question 11.17 can be connected to patterns such as those obtained in X-ray diffraction studies of crystal powders. Student discussion and interpretation of the connection greatly enhances the richness of this context and exposes the students to a valuable exercise in qualitative thinking.

11.18 Suppose a monochromatic point source of light P is placed very close to a good plane mirror as shown in the diagram. The result is an interference pattern in the region above the mirror, the effect being known as “Lloyd’s Mirror.”



- (a) Explain the origin of the interference pattern, making a sketch of wave fronts to show how the pattern is developed in the plane of the diagram. What is the independent variable that determines the angular spread of the pattern? How is this situation connected with the two-slit interference pattern you have already studied? Explain.
- (b) Making use of results you have already derived in class or textbook, write equations for the angular location of constructive and destructive interference loci in the plane of diagram. The angular positions should be expressed in terms of the relevant independent variable you defined in part (a).
- (c) Describe how you might set up an analogous situation in a ripple tank.

Chapter 12

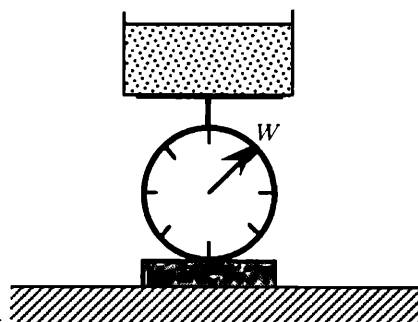
Fluids and Thermal Phenomena

12.1 A circular cylinder, open to the atmosphere, contains 2500 cm^3 of a liquid having a mass of 0.87 g in each cubic centimeter. The atmospheric pressure is 1.00 bar . (Be sure to explain your reasoning in each of the following calculations.)

- (a) Calculate the total mass of the liquid in SI units.
- (b) Calculate the total weight of the liquid in SI units.
- (c) If the height of the liquid in the cylinder is 120 cm ,
 - (1) Calculate the gauge pressure p_{1g} at the bottom of the cylinder.
 - (2) Calculate the absolute pressure p_{abs} at the bottom of the cylinder.
- (d) If all the liquid is poured into a second cylinder with a diameter 1.80 times larger than that of the first cylinder, calculate the ratio of the gauge pressure p_{2g} on the bottom of the second cylinder to the gauge pressure p_{1g} previously exerted on the bottom of the first cylinder. (Solve this problem by ratio reasoning, not by substitution in formulas.)
- (e) Is the ratio of absolute pressures at the bottoms of the cylinders in part (d) the same as the ratio of the gauge pressures? Why or why not? Explain your reasoning.

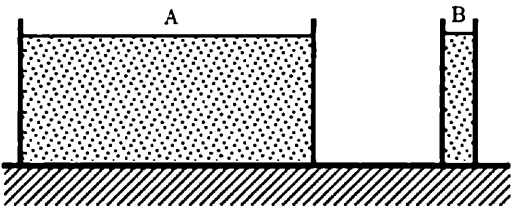
12.2 A container of water rests on a platform scale as shown, and the needle gives the reading W for the weight being measured.

- (a) How does the reading W compare with the force exerted by the water on the bottom of its container plus the weight of the container, i.e., is W equal to, greater than, or less than this combined force? Explain your reasoning.



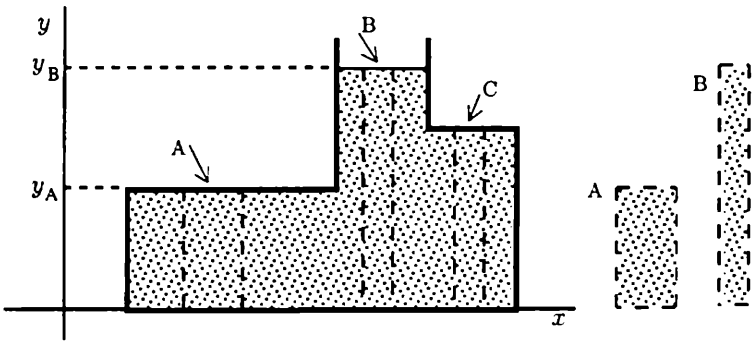
- (b) Does the reading W include the weight of the column of air above the container? Why or why not?
- (c) Would you expect the reading W to change significantly if the atmospheric pressure were decreased? Why or why not? If you expect it to change, would it increase or decrease? Explain your reasoning.

12.3 Consider two rectangular tanks of water: The side walls (left and right) of the tanks have exactly the same area, but the front and back walls do not because the lengths of the tanks are quite different. Both tanks are filled to the same depth and obviously contain very different total weights of water.



- (a) How does the total outward force on the left and right side walls of container A compare with the total outward force on the left and right side walls of container B? Are the forces equal or is one greater than the other? Explain your reasoning.
- (b) How do the total outward forces on the front and back walls of container A compare with the corresponding forces in container B? Explain your reasoning.

12.4 The diagram represents a container filled with liquid of density ρ . In our imagination we shall “extract” from the interior of the liquid the two columns labeled A and B. Let us denote the cross-sectional (or base) areas of the two columns by the symbols ΔS_A and ΔS_B and their heights by y_A and y_B , respectively. We denote the atmospheric pressure by p_o .



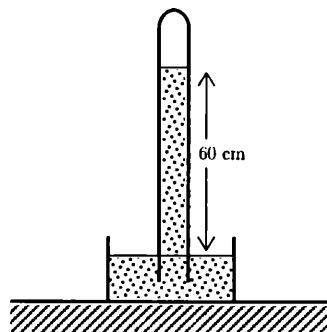
- (a) Draw free body force diagrams for columns A and B on the right, showing all the *vertical* forces being exerted on each. (To avoid cluttering the diagram, do not try to show horizontal forces on the columns even though we know such forces are being exerted.) Describe each vertical force in words. (Verbal description consists of saying what object exerts this force on what.)

- (b) In terms of the symbols defined above, what is the algebraic expression for the total force acting on the top surface of column B? On the top surface of column A? Explain your reasoning.
- (c) Including expressions for the weight of each column, what must be the expression for the total force acting on the bottom surface of column B? On the bottom surface of column A? Explain your reasoning.
- (d) Do your expressions lead to the result that the pressure is the same at the base of each column? If not, review your work because you must have made an error in reasoning.
- (e) Is the absolute pressure at the bottom of this container equal to the pressure of the atmosphere plus the total weight of the liquid divided by the total area of the base? Why or why not? Under what circumstances, if any, is the absolute pressure at the bottom of a container of liquid equal the pressure of the atmosphere plus the total weight of the liquid divided by the total area of the base?
- (f) How does the situation at the top of column C compare with that at the top of A and B? Explain your reasoning.
- (g) Suppose a hole is made in the wall of the container at the very top of column B. What will happen? Explain your reasoning. Suppose a hole is made at the top of column A?

12.5 An apocryphal story goes as follows: A large van stops at a traffic light, and the driver leans out and pounds on the wall with a baseball bat. A puzzled driver in an adjacent lane asks what is going on. The van driver replies, "This is only a five-ton van, and I am carrying eight tons of canaries, so I have to keep at least three tons of them flying."

Assess the van driver's physics. Can he get away with this strategy? Why or why not? What must actually be happening inside the van when the canaries are flying? What is the force on the floor of the van regardless of whether the canaries are roosting or flying? Explain your reasoning.

12.6 Consider this barometer-like arrangement. Atmospheric pressure has its normal value corresponding to 76 cm of mercury, but the column of mercury in the tube stands at a height of $y = 60$ cm instead of 76 cm above the surface of the mercury in the reservoir, and there is a space between the top of the mercury column and the top of the tube.



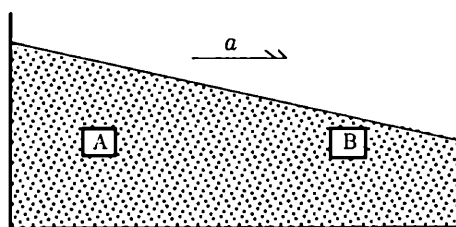
- (a) What do you infer about the contents of the space in the tube above the top of the mercury column compared with the contents of this space if the column height were 76 cm? Explain your reasoning.

- (b) How does the pressure in this space compare with atmospheric pressure: Is it equal to, greater than, or less than atmospheric pressure? Calculate the ratio of any pressure in the space to atmospheric pressure. Explain your reasoning.
- (c) What would happen to the height y of the mercury column if the outer container were very much deeper and more mercury were poured into the reservoir, raising its height in the outer container? Would y increase, decrease, or remain unchanged? What would simultaneously happen to the pressure in the space above the top of the column? Contrast this result with what would happen if the space above the column were a perfect vacuum. Explain your reasoning.

12.7 Let us imagine an iceberg as a neatly rectangular block of ice floating in seawater. The density of sea ice, with its inclusions of air bubbles, snow, and drops of liquid water, is about 0.89 g/cm^3 . The density of seawater is about 1.03 g/cm^3 .

- (a) Draw a force diagram of the floating block, representing the distributed force exerted by the surrounding water by a single total force. Describe each force in words.
- (b) Stating the appropriate governing principle or principles and explaining each step, calculate the fraction of the height H of the block that will stick up above the water surface.
- (c) Consider a second block of sea ice having the same density as the one above but scaled down in size, with all three length dimensions being half those of the original. What fraction of the height of this second block will protrude above the water surface? Be sure to explain your reasoning.

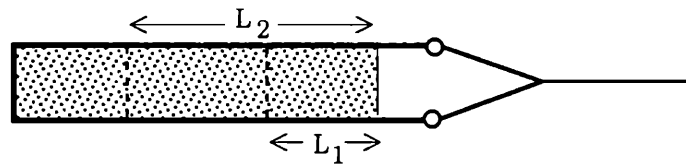
12.8 If a tank of water is accelerated to the right (in a car or on a cart), the water shifts to the left, and the surface takes on a slope downward to the right as shown in the following diagram. A constant downward slope is maintained as long as the acceleration is constant.



- (a) Every particle within the body of water is being accelerated to the right, along with the entire system. Draw separate free body force diagrams for the parcel of fluid in box A and the parcel of fluid in box B, showing larger forces with longer arrows and equal forces with arrows of equal length. (Use single arrows for forces on the sides of each box, even though the force is actually continuously distributed; this is what we usually do, for example, with normal, gravitational, and frictional forces, which are also continuously distributed.) Explain any reasoning behind the drawing of unequal forces.

- (b) Use your force diagrams to explain why each of the parcels of fluid has the same horizontal acceleration and zero vertical acceleration even though they are located at unequal depths within the fluid.
- (c) Suppose a pendulum bob is suspended from the roof of the same accelerating car. How would you expect the angle θ between the pendulum string and the vertical to compare with the angle between the water surface and the horizontal? Explain your reasoning.

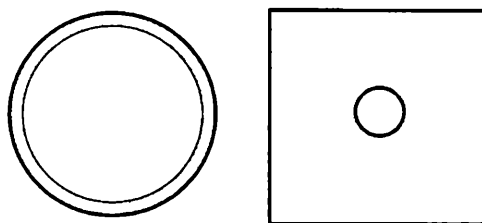
12.9 A cylinder of water fastened to a string is whirled around at constant angular velocity in a circle lying in very nearly a horizontal plane. The diagram shows an instantaneous view of the cylinder from the side when its axis happens to lie in the plane defined by the paper.



- (a) Draw two separate free body force diagrams for the two columns of water of lengths L_1 and L_2 , respectively. (Draw longer arrows for larger forces and arrows of equal length for forces of equal magnitude.)
- (b) Explain how your diagrams account for the centripetal forces necessary to keep all the water in the tube moving in a circle.
- (c) Argue that what you have said in parts (a) and (b) means that there must be a pressure gradient (continuously increasing pressure) from right to left in the fluid. Compare this with the pressure gradient that exists in the fluid when the cylinder simply stands upright in the laboratory.
- (d) Suppose a small cork is floated in the cylinder before the system is put into revolution. Where will the cork finally be located in the fluid in the steadily revolving cylinder? Explain your reasoning.
- (e) Suppose the cylinder is whirled around in a vertical instead of a horizontal circle. Again by drawing force diagrams, examine the forces acting on the columns L_1 and L_2 at the instants the cylinder is at the top of the circle, at the bottom of the circle, and halfway in between (i.e., at the instant the string is horizontal). Describe the differences and similarities between each of these three cases on the one hand and the situation in the horizontal circle on the other.
- (f) When the cylinder is revolved in a vertical circle, under what circumstances will the liquid remain in the cylinder when the latter is at the top of the circle and under what circumstances will the liquid tend to fall out? Explain your reasoning with the help of appropriate force diagrams.

12.10 Consider the following objects: A thin brass ring and a brass plate (of the same composition as the ring) with a hole in it. The two objects start at the same

temperature, and the temperature of each is then increased slowly and uniformly (i.e., the temperature does not vary *within* each object).



Discuss what will happen to the dimensions of each object, i.e., to the inner and outer diameters of the ring, and to the lengths of the outer edges of the plate and the diameter of the hole in the plate. We are not concerned with numbers but only with directions of change: Increase, decrease, or no change. Explain your reasoning carefully, not only to defend your conclusions, but also to be able to refute conclusions that differ from yours. Give especially careful attention to what happens to the inner diameter of the ring and to the diameter of the hole in the plate.

12.11 The coefficient of thermal expansion β of liquid water varies appreciably with temperature at fixed atmospheric pressure. Experimental measurements of β between 10 °C and 20 °C, at atmospheric pressure, are fairly well represented by the following empirical relation [where $\beta(t)$ indicates that β is a function of the Celsius temperature t]:

$$\beta(t) = 87.9(10^{-6}) + 11.9(10^{-6})(t - 10)$$

- Define the term “empirical” in the foregoing context.
- Define the term “coefficient of thermal expansion.”
- Making appropriate use of the calculus and explaining each step of reasoning, calculate the change in volume that takes place when 1500 cm³ of liquid water increase in temperature from 10 °C to 20 °C at constant atmospheric pressure.
- Sketch a graph of $\beta(t)$ versus t , and interpret the change in volume you have calculated in part (b) as a quantity related to some property or feature of this graph. How is this graph related to a graph of volume versus temperature?
- In terms of what you have said and done in parts (b) and (c), explain, as though you were dealing with an uncertain fellow student, why it is incorrect to say that the volume change can be found by simply multiplying out $1500 \times (87.9) \times (10^{-6}) \times (10)$. Connect your explanation with the graphs sketched in part (d) as well as with the arithmetic involved.

12.12 Although liquid-in-glass thermometers are usually made of glass that expands very little on increase in temperature, some expansion of the glass does occur in most instances, especially for fairly large temperature changes.

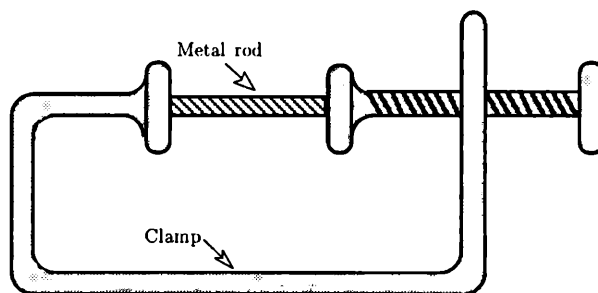
- As the glass expands, what happens to the volume of the inner space occupied by the liquid? Does it increase or decrease? Explain your reasoning.
- How do you explain the fact that the liquid does rise in the thermometer with increasing temperature even though the glass expands? Cite your evidence and explain your reasoning carefully.

- (c) Do you expect the scale of graduations on a thermometer to be perfectly uniform over the entire length of the instrument? Why or why not?

12.13 (1) It is an observed fact that liquids and solids can be compressed (i.e., made smaller by application of external forces), but it takes very, very large forces to produce very small changes in volume (or length). (2) You are certainly aware that bottles (or water pipes) burst when water that fills them freezes.

- (a) What information does fact (2) give us about the behavior of the volume of a parcel of water when it freezes? Explain your reasoning.
- (b) What connection do you see between fact (1) and the bursting described in fact (2)? Explain your reasoning.

12.14 A metal rod is held firmly within the jaws of an iron clamp. A blowtorch is now directed at the metal rod (not the clamp), and the temperature of the rod is increased without appreciable change of the temperature of the clamp.



What is likely to happen to the clamp as the temperature of the rod increases? Explain your reasoning. How is the effect you are describing connected with our awareness that although solids can be compressed (made smaller in size) by application of external forces, very large forces are required to produce very small changes in dimensions?

12.15 It is a matter of everyday experience that when objects at different temperatures are brought in contact, an “interaction” takes place. The temperature of each object changes until a stopping point (equilibrium) is reached. We call this a “thermal” interaction. (We recognize, in the world around us, many other kinds of interaction in which objects, or groups of objects “do something” to each other. We speak of mechanical, gravitational, electrical, magnetic, chemical interactions, etc.) Let us reexamine our everyday experiences to pull out the regularities in thermal interactions that remain hidden unless we think about them.

- (a) We have two containers of water, each with its own thermometer. What happens if we bring the two, one at higher and one at lower temperature, in contact with each other, either by simply mixing the waters or by good, direct contact between the containers? (If we imagine doing such an experiment by direct contact between containers, it is well to imagine separating the two containers from contact with air in the room by keeping them in a good, thick box.) In which direction (increase or decrease) does each thermometer change when

contact begins? At what point as far as the thermometers are concerned is equilibrium reached? When, that is, do changes cease? If the two amounts of water are different, which ends up with the bigger temperature change when equilibrium is attained? If you made one of the water quantities very, very much larger than the other, where would the temperatures end up? How would the temperature change of the huge body compare with that of the tiny one? Is the temperature change of the large body ever actually zero?

- (b) What is the end point of temperature changes for a container of hot water put out in a room that is initially at room temperature (20°C)? What bodies undergo temperature changes in the interaction? What would happen to the temperature of the air in the room if the water container were very large? Answer the same questions for the case in which the container of water is very cold instead of very hot. How do these situations relate to the ones you discussed in part (a)? How is the temperature of a room maintained at a comfortable level during cold weather?
- (c) Placing materials such as wood, glass, or fiber between thermally interacting objects greatly slows down the temperature changes but does not make them stop at unequal temperatures. We speak of such layers of material as providing “thermal insulation.” Explain, in the light of your discussions in parts (a) and (b), the point of thermally insulating a house and the point of using containers that tend to keep cold drinks cold.
- (d) Have you ever seen a case in which a colder object gets colder and a hot one hotter on contact? Or a case in which two objects start at the same temperature and one spontaneously increases in temperature while the other decreases? Why would it be exceedingly desirable to be able to make a kettle of water boil by bringing it into contact with an object at the same initial temperature? It is perfectly possible to *imagine* such changes, but do you believe them to be possible? Why or why not?
- (e) What generalizations about thermal phenomena can you now put together in the light of visualizing and relating the everyday experiences we have been describing?

12.16 We tend to use the word “heat” very casually in everyday speech, and many individuals use the words “temperature” and “heat” synonymously; i.e., they do not distinguish the terms as having very different meanings. In physics, “heat” is a technical term and denotes an abstract concept that is defined when the need is recognized through experience. Let us, in this question, retrace some of this experience. We start only with the idea of “temperature” as the reading on the familiar thermometer. When we examine some common thermal interactions, we begin to find that the thermometer does not tell us the whole story concerning the interactions (if it did, we would not need to invent an additional concept.)

- (a) Consider three identical quantities of hot water, all at exactly the same initial temperature. We put them out in a room at a fixed, lower temperature, but we use three different containers. The first sample is contained in an ordinary glass with no wrapping. The second sample is placed in a similar glass, but the glass is wrapped with a thick layer of plastic material such as styrofoam. The

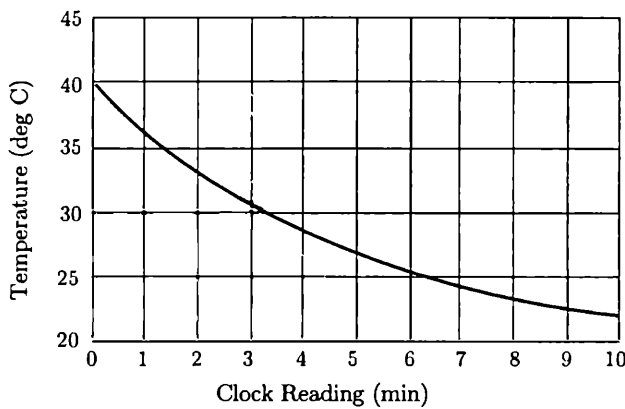
third sample is placed in a high quality thermos bottle. Note carefully what your experience tells you: All three samples start at exactly the same initial high temperature and end up at exactly the same (lower) room temperature. So there is no difference between the initial temperatures and no difference between the final temperature readings when equilibrium is attained. There is, however, a very significant difference between the intermediate histories of the three samples. Describe what actually happens. How do the three interactions differ even though the initial and final temperatures do not differ?

- (b) Suppose we start with two very different quantities of water (one larger and one smaller) in separate beakers, initially at exactly the same temperature. We place the beakers on identical heaters and proceed to increase the temperatures in each beaker to the same final value. The temperatures in both samples are the same to begin with and the same at the end; there is no difference among the thermometer readings. Yet there are significant differences between the two interactions. Describe the differences in your own words, being explicit about the time intervals involved and the utilization of fuel or electricity in the heating.
- (c) Suppose we start with two equal amounts of water, one sample at room temperature (20°C) and the other at, say, 80°C . We immediately mix the two samples. At what temperature do you expect equilibrium to be attained (in absence of significant interaction with the surrounding air in the room)? Explain your reasoning. Suppose the higher temperature sample is made considerably smaller than the lower temperature one? Where does experience say the equilibrium temperature ends up? In what way do the thermometer readings alone fail to tell us the whole story of interaction?
- (d) If we have a solid material in a test tube that is surrounded by a bath that is high enough in temperature, we might bring the material to its melting point (provided we stay below the melting point of the containers). It is an experimental fact that, if the material in the test tube does melt, the thermometer immersed in this material undergoes no change while the solid material is disappearing and begins to show an increase in temperature only after all the material is melted. Explain in your own words why this experimental observation is a dramatic illustration of the fact that thermometer readings alone do not tell us the whole story of thermal interaction.
- (e) Explain in your own words why the foregoing examinations of familiar thermal phenomena lead us to invent a new concept, the concept to which we give the name "transfer of heat."

12.17 Liquid water has a nonzero vapor pressure at all temperatures, as confirmed by the fact that it evaporates at any temperature if left standing. If water is evaporating at all temperatures, what is so special about 100°C ? What is it that happens when boiling occurs that is different from what is happening when water evaporates at room temperature?

12.18 If we put some hot water (at initial temperature T_1 at 40°C) in a beaker and place the beaker in a room where the air temperature has a lower value T_2 at 20°C , we know that the temperature of the water drops as time goes by, ending up at the same temperature as the air. If we keep track of the water temperature as a function of

clock reading (keeping the water well stirred) and graph the data, we obtain a graph roughly like that depicted in the following figure. Such a graph is called a “cooling curve.”



- (a) Suppose you started with the same amount of water at T_1 but wrapped the beaker in styrofoam (or set the beaker in a tightly fitting styrofoam cup.) You now measure the new cooling curve. Sketch what you would expect the curve to look like in comparison with the one sketched in the diagram.
- (b) Make similar sketches for the cooling curves you would expect if (1) you put the water in a good thermos bottle and (2) you put the water in a thin metal container.
- (c) What is the connection between the sequence of cooling curves you have sketched and the point and purpose of insulating a house in the wintertime? Explain your reasoning.

12.19 Joseph Black (1728-1799) was a Scottish physician who, at various times, was professor of medicine and chemistry at the universities of Glasgow and Edinburgh. [Among his students was James Watt (1736-1819), whose improvements and inventions in the design of steam engines played an important role in the advent of the Industrial Revolution in the 19th century.] Black was a leader among the researchers who, around the middle of the 18th century, refined and sharpened the concepts of “temperature” and “heat.” Black was probably the first to recognize the significance of the thermal interaction taking place during melting and freezing, when the process of change of phase of a single substance takes place without change in temperature. About this phenomenon, he wrote:

Melting has been universally considered as produced by the addition of a very small quantity of heat to a solid body, once it has been warmed up to its melting point; and the return of the liquid to the solid state as depending on a very small diminution [of heat]. . . It was believed that this small addition of heat during melting was needed to produce a small rise in temperature as indicated by a thermometer.

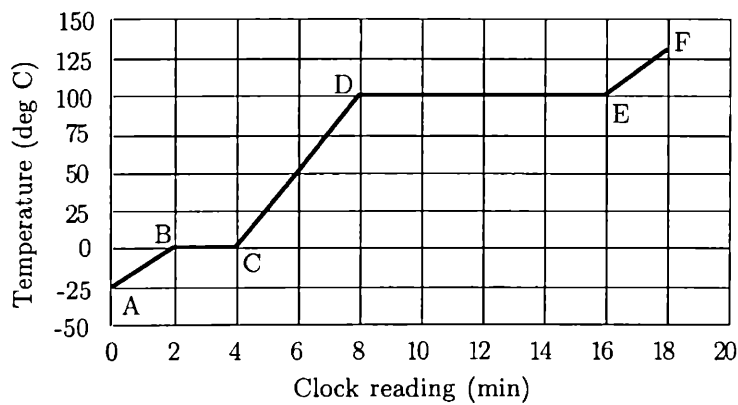
The opinion I formed is as follows. When ice or any other solid substance is melted . . . a large quantity of heat enters into it . . . without making

it apparently warmer as tried by [a thermometer]. . . . I affirm that this large addition of heat [without change in temperature] is the principal and most immediate cause of the liquefaction induced. . . .

If the common opinion had been well founded—if the complete change of ice and snow into water required only the addition of a very small quantity of heat—the mass, though of considerable size, ought to be all melted within a few minutes or seconds by the heat incessantly communicated from the surrounding air. Were this really the case, the consequences of it would be dreadful . . . for even as things are at present, the melting of large amounts of snow and ice occasions violent torrents and great inundations in the cold countries. . . . But were ice and snow to melt suddenly, as they would if the former opinion of the action of heat . . . were well founded, the torrents and inundations would be incomparably more irresistible and dreadful. . . . This sudden liquefaction does not actually happen. The masses of ice and snow require a long time to melt. . . .

Discuss Black’s argument, translating it into your own words and connecting it with your own experiences. Why do you think Black felt compelled to present this argument to his contemporaries? (Note that what he was doing was creating the then new concept of “latent heat.”)

12.20 Following is a schematic (highly idealized) presentation of a temperature versus time graph for an experiment in which a sample of solid ice is taken out of a deep freeze at around -25°C and heated uniformly in a closed container to a final temperature in excess of 100°C . The system is “closed”; i.e., no material enters or escapes from the system. (The presentation is schematic in the sense that real experimental data would show scatter and uncertainty not represented in the graph.)



The following questions pertain to what is happening to the sample during the time intervals indicated by the letters on the graph. For example, during the time interval AB the temperature of the solid ice is increasing with time.

- (a) What is happening to the sample during the interval BC?
- (b) What is happening to the sample during the interval CD?

- (c) What is happening to the sample during the interval DE?
- (d) What is happening to the sample during the interval EF?
- (e) Under what circumstances is the history denoted by EF possible? What would happen after instant E if the container were open to the atmosphere rather than closed?

12.21 Joule, in his classic “paddle wheel” experiment (made to determine how much heat, measured in calories, corresponded to a given amount of work measured in mechanical units), used the lowering of a weight to drive a paddle that churned water in a thermally well-insulated container and thus raised the temperature of the water slightly as the mechanical effects were dissipated. (In the following, we shall use modern units, rather than Joule’s old English units.)

Suppose that after a block with a mass of 1.80 kg is lowered 16 times through a height of 11.0 m, the 3.00 kg of churned water in the thermally well-insulated bucket exhibits a temperature rise of $0.23\text{ }^{\circ}\text{C}$. Take the mass of the bucket to be 1200 g, and the specific heat of its material to be $0.107\text{ cal/(g)} (^{\circ}\text{C})$.

- (a) Starting with the block in its elevated position, describe in your own words all the energy changes and transformations that take place as the block is lowered.
- (b) From the data given above, calculate how many joules of dissipated mechanical work have the same thermal effect on the system as would the transfer of one calorie of heat. (Note that no heat was transferred to the water in this experiment; mechanical work was dissipated.)
- (c) Suppose the water and the bucket start off at the same temperature as the surrounding air. What effect would a slight thermal interaction with the surroundings (through the insulation on the bucket) have on the result of your calculation? That is, would the interaction tend to make the calculated result higher or lower than the “correct” value? Explain your reasoning carefully on the basis of the equation you set up in making the calculation.

12.22 A plastic tube capped at both ends contains a quantity of lead shot. A common laboratory experiment for determining the number of joules corresponding to one calorie consists of inverting the tube quickly (so that the shot are carried to the top), allowing the shot to fall to the bottom, and repeating the process a number of times. The temperature change of the shot is measured after a recorded number of such falls.

- (a) Describe in your own words the sequence of energy changes and transformations that takes place in this experiment. Why does this experiment make it possible to calculate the number of joules corresponding to one calorie?
- (b) In an experiment in which the length of the tube was such that the center of mass of the shot fell through a distance of 1.0 m, the initial temperature of the shot was 22.0°C , and the final temperature, after 50 inversions of the tube, was 25.6°C . Look up any other data you need in an appropriate table, and calculate what this experiment yields for the number of joules corresponding to one calorie. Be careful about the justified number of significant figures in your final result. Explain all steps of reasoning, and be explicit about any idealizations or approximations introduced in your analysis.

- (c) If your result in part (b) differs from the accepted value given in the text, does the deviation make sense? That is, is it in a direction you would expect from the way the experiment is carried out and from the idealizations involved in your calculation? Explain your reasoning.
- (d) Why use a plastic tube? Why not a metal tube, for example? Why use lead as the falling material? Why not iron, or copper, or aluminum? (Look up the numerical value of the relevant property of these materials in the appropriate table and make use of this information in responding to the question.) Explain your answers carefully, making reference to the terms that would be affected in the equation you set up in part (b) and inferring the consequences to the accuracy of the experiment. Might the experiment be significantly improved by using a longer tube? Why or why not?

12.23 Several small pieces of copper, having a total (combined) mass of 350 g are placed in liquid nitrogen and, when removed, are at a temperature of -180°C . The pieces of copper are quickly transferred to a calorimeter containing 420 g of water at $+9.0^{\circ}\text{C}$. The calorimeter cup is made of brass [specific heat $0.092 \text{ cal}/(\text{g})(^{\circ}\text{C})$] and has a mass of 200 g. The temperature of the room in which the experiment is conducted is $+20^{\circ}\text{C}$.

- (a) Describe *qualitatively* what happens in the way of heat transfers and temperature changes after the copper has been placed in the calorimeter. Describe the conservation relation that governs the phenomena taking place, and indicate the idealization we make in applying this relation to the interaction between the copper and the calorimeter. What role does the concept of “closed system” play in the idealizations you invoke?
- (b) Now proceed to predict, numerically, the final equilibrium state that is attained within the calorimeter. Explain your reasoning as you set up quantitative expressions and interpret, in words, the physical meaning of each separate term that is present in the equation you end up with. (Hint: You would be wise to make a preliminary, rough calculation to estimate the final physical condition the system attains and to identify the relevant unknown, or unknowns. Otherwise you may find yourself putting numbers into expressions that turn out to be irrelevant.)
- (c) Noting that it is, in fact, never possible to attain a perfectly closed system in these circumstances, analyze how interaction with the surroundings will, in this instance, affect the final result: Will the calculated value (of whatever you calculated) be greater than, equal to, or less than the “correct” value that would have been obtained in the ideal situation?

12.24 A refrigerator is left running with its door open in a tightly closed, well-insulated room. As time goes by, what happens to the temperature of the room: Does it increase, decrease, or remain unchanged? Explain your reasoning.

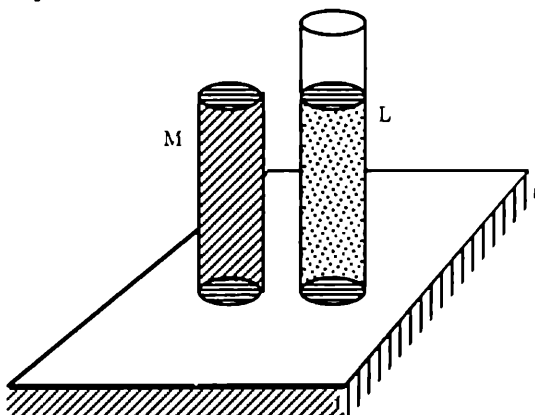
12.25 It is a familiar fact that when electric current is passed through a wire (as in the electric stove, the toaster, the clothes iron, the light bulb, the experiment in the laboratory), the temperature of the wire increases to some point at which

the increase ceases. How do you account for the fact that the temperature does not keep increasing indefinitely? What interactions, processes, and effects come into play to stop the increase? What role does the concept of “equilibrium” play in this phenomenon? Explain in as much detail as you can.

12.26 A bicycle pump is full of air at atmospheric pressure; i.e., the piston is as far out in the cylinder as it will go. The stroke of the piston is 14 inches: i.e., the end of the cylinder is 14 inches away from the face of the piston in the initial state. How far must the piston be pushed in before air will begin to enter the tire in which the gauge pressure is 40 lb/in²? State all your assumptions and explain your reasoning. Be careful about the distinction between absolute and gauge pressures and the role they play in your calculation. What would be the effect on your numerical estimate (i.e., Will your estimate be higher or lower than the actual value?) if there were some leakage of air past the piston as compression took place? What assumptions have you made about temperature changes, and how do you justify them? What would be the effect on your numerical estimate of any realistic departure from whatever assumption you have made about temperature changes?

Note to the student: In the following multiple choice questions, circle the letters designating those statements that are correct. *Any number* of statements may be correct in each question, *not* just one. You must examine each statement on its merits.

12.27 A cylindrical metal rod M stands on the table. Also standing on the table is a cylinder L containing a liquid. The cylinder is made of a glass called “invar,” which does not expand or contract appreciably on change in temperature. Suppose the temperature of the room is increased with corresponding uniform increases in temperature of both systems L and M. Both the metal and the liquid undergo expansion with increase in temperature.



- (a) The density both of rod M and of the liquid in L will decrease as the temperature increases.
- (b) To calculate the pressure at the bottom of the liquid in L at the new temperature, one would make use of the relation $pV = nRT$, being careful to express the temperature in degrees kelvin.

- (c) To calculate the *change* in pressure at the bottom of the rod and at the bottom of the liquid, one would make use of the relation $\Delta p = \rho g \Delta h$, where Δh represents the increase in height that takes place on expansion in each system.
- (d) Before the temperature change, the pressure at the bottom of each column would be the same if the heights were the same.
- (e) The pressure at the bottom of M will increase as the rod gets longer.
- (f) The pressure at the bottom of M will decrease slightly as the rod expands.
- (g) The pressure at the bottom of the liquid in L will increase as the temperature increases and the height of the liquid increases.
- (h) The pressure at the bottom of the liquid in L will remain unchanged as the temperature changes.

12.28 A flat block of paraffin floats on water with 4.0 mm of its thickness projecting above the water surface. A hole, about 2 cm in diameter, is drilled through the center of the block, which is then put back in the water. The block of paraffin now

- (a) sinks because of the hole that has been drilled.
- (b) floats with its upper surface level with the water surface.
- (c) floats with somewhat less than 4.0 mm projecting above the water surface.
- (d) floats with somewhat more than 4.0 mm projecting above the water surface.
- (e) floats exactly as it did before the hole was drilled.
- (f) turns up on edge.

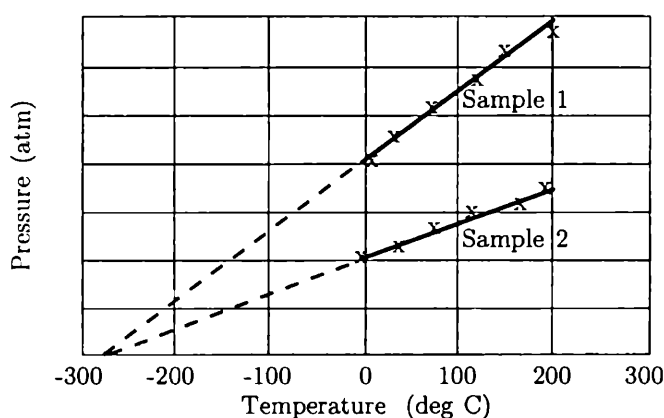
12.29 As water boils at 100°C at atmospheric pressure, bubbles of gas form in the interior of the liquid and rise to the surface. The bubbles contain primarily

- (a) carbon dioxide.
- (b) air.
- (c) vacuum.
- (d) a mixture of the separate gases hydrogen and oxygen.
- (e) None of the above.

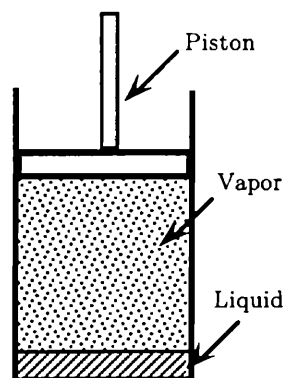
12.30 Two different samples of the same gas are placed in rigid containers having the same volume. Measurements of pressure and temperature made on the two samples are plotted in the following graph. The data, which show relatively small scatter, are extrapolated by straight lines to zero pressure.

- (a) Both samples are exhibiting a close approximation to ideal gas behavior.
- (b) Compared with sample 2, sample 1 must contain a larger mass of gas.

- (c) Samples 1 and 2 must contain the same mass of gas.
- (d) At any given pressure, sample 2 always has the higher temperature.
- (e) Some of sample 2 must have been continually leaking away during the measurements.
- (f) The two samples disagree with respect to definition of the absolute temperature scale.
- (g) The behavior of the two samples points toward definition of the absolute temperature scale.
- (h) None of the above.



12.31 The cylinder contains a small amount of liquid water at the bottom. Above the water, the pure water vapor is confined by the piston, which can be raised or lowered. There is no air present; the gas phase contains water vapor only. The piston walls are made of good heat-conducting material, ensuring that any temperature change within the cylinder is quickly wiped out by conduction to or from the air in the room if the piston is moved quite slowly. Thus the temperature within the cylinder remains very nearly constant. Leakage of gas past the piston is to be considered negligible. As the piston is moved slowly downward:



- (a) the pressure in the cylinder remains unchanged.
- (b) the amount of liquid increases while the amount of vapor decreases.
- (c) heat flows out of the system to the surrounding air.
- (d) none of the above.

Chapter 13

Kinetic Theory

13.1 Once we have accepted the picture of discreteness in the structure of matter (the existence of atoms and molecules as separate entities in the architecture of material substances), we proceed to visualize gases as consisting of atoms or molecules that are (1) in continual motion, colliding with each other and with the walls of the container, and (2) on the average, very far apart relative to their own diameters.

- (a) What are the justifications for this mental picture or model? In answering this question, draw on what you see in the actual behavior of gases and on what you know about their macroscopic physical properties.
- (b) Why is this identical model not applicable to liquids or solids? What properties of liquids and solids indicate that their atoms or molecules are not far apart relative to their diameters?

13.2 In the model for gases referred to in Question 13.1, what is the justification for assuming collisions of the molecules with each other and with the walls of the container to be perfectly elastic? In answering this question, you should appeal to what we actually observe in the behavior of gases and what you would *expect* to observe if the collisions were *not* perfectly elastic. In other words, you must visualize, in the abstract, things that do not actually happen. (It is often the case in science that visualizing something that does *not* happen is just as important, and just as conducive to understanding, as knowing or visualizing what does happen.)

13.3 In the model referred to in Question 13.1, how do you interpret the macroscopic property of “density” of a gas in terms of molecular numbers? In visualizing the behavior of the continually moving atoms or molecules, how do you account for the uniformity of the density throughout any container? (When a system is uniform in all directions, we say it is “isotropic.”)

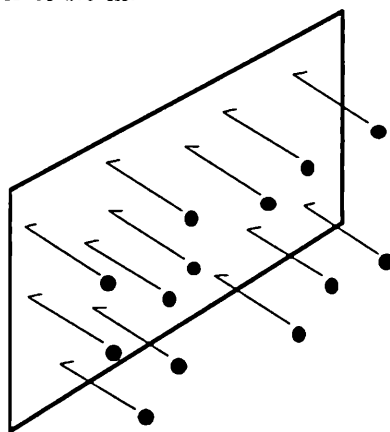
13.4 In the model referred to in Question 13.1, how do you interpret, in terms of molecular behavior, the macroscopic pressure that any gas exerts on the walls of its container? In other words, what molecular behavior generates the pressure? How do you account for the steadiness of the pressure reading? In other words, why doesn’t the needle of a pressure gauge keep jiggling randomly as individual molecules keep bumping the sensitive area of the gauge? How do you account for the uniformity of the pressure over the entire surface of the container? Suppose you had access to

an extremely sensitive pressure gauge with an extremely small surface area. How would you expect the readings of this gauge to behave, especially as you decreased the amount of gas within the container? Explain your reasoning.

13.5 It is an observed fact that if we increase the temperature of a gas and keep the *volume constant*, the pressure of the gas increases. Suppose we now accept the idea that the gas consists of discrete particles or molecules. Argue that the observation we have described leads us to infer that increased temperature must be associated with increased molecular velocity. Explain why the constant volume restriction is a necessary part of the argument.

13.6 Balls of putty, each having a mass of 2.5 g, are projected in a steady stream against a wall as shown. The balls all have a velocity of 130 m/s perpendicular to the wall, and they stick to the wall on striking it. Over an interval of 5.0 s, 1600 balls strike a circular area of the wall having a diameter of 1.0 m.

- (a) Explaining your steps of reasoning, calculate the average pressure exerted on the wall by the rain of balls of putty.
- (b) Compare your result in part (a) with the magnitude of normal atmospheric pressure: Is it large or small compared with atmospheric pressure? What significance do you see in the difference?



13.7 Suppose we start with a container of gas that has had a burner applied to one end so that the temperature of the gas at that end has been increased well above the temperature in the rest of the container. (Under such circumstances, we say that we have created a temperature *gradient* in the container.) We now remove the burner. Describe, in terms of molecular behavior and molecular kinetic energy, what happens as the gas returns to thermal equilibrium and the temperature gradient disappears. Explain the connection between the description you have just given on the microscopic level with what we previously called “transfer of heat” on the macroscopic level.

13.8 When we compress a gas by displacing a piston in a cylinder (as in a tire pump) against the opposing pressure of the gas, we must do work *on* the system (gas) to effect this compression. Describe in your own words what happens in the way of energy transformations in this process. There are several different cases to be considered:

- (a) Consider the case in which the compression is performed fairly quickly and the cylinder is well insulated thermally from its surroundings. [In this process there is negligible heat transfer between our system (the gas, piston, and cylinder) and the surrounding air. Such a process is called “adiabatic.”] In a macroscopic sense, what happens to the work we have done on the system? Into what other form or forms of energy has it been transformed? How do you describe and

interpret the same process on the microscopic level? (Note: It is an observed fact that in adiabatic compression, the temperature of a gas always increases. In some rare instances involving other substances, e.g., water between 0 °C and 4 °C, temperature decreases on adiabatic compression.)

- (b) If the cylinder is not thermally insulated and we compress the gas slowly, we can effect a compression in which the temperature of the gas remains essentially constant. Such a process is called “isothermal.” Describe the energy transformations taking place in this process, first in macroscopic, and then in microscopic, terms. What happens to the work done on the system?

13.9 Two glasses hold liquid water at room temperature: (1) One glass is open to the surrounding air, and (2) the other glass is tightly covered but has an air space above the water it contains. Imagine an initial condition in which we have just placed water in each container and there are, as yet, no water molecules in the air above the liquid.

- (a) With sketches and verbal explanation, describe, in terms of the random motion of molecules of both water and air, what happens as time goes by after the initial condition specified above. What is the difference between cases (1) and (2)? How do you account for the fact that all the liquid eventually disappears in case (1) while the liquid does not disappear in case (2)? (As you have surely recognized, the technical name for the disappearance of the liquid under such circumstances is “evaporation.”)
- (b) In what sense is the term “equilibrium” relevant to the situation described in case (2)? Is a condition of equilibrium one in which molecular motion and migration of molecules ceases? Is the term “equilibrium” relevant to case (1)? Explain your answers in each instance.
- (c) Compare any heat transfer that takes place between the air in the room and the liquid in cases (1) and (2): If any heat transfer does take place, what is the direction of the transfer? If no heat transfer takes place, say so explicitly. In either case, explain your reasoning.

13.10 A quantity of solid sugar is placed in a beaker containing liquid water. The sugar proceeds to dissolve in the water.

- (a) With sketches and verbal explanation, describe what happens at the molecular level as the sugar dissolves, including a description of how the sugar molecules spread out (diffuse) through the water beyond the immediate vicinity of the solid sugar itself.
- (b) Using a molecular level description similar to that given in (a), compare what happens when only a small amount of sugar is placed in the beaker, and all the solid disappears, with the situation that develops when a large chunk of sugar is placed in the beaker and the disappearance of the solid ceases at some point.
- (c) In the light of the descriptions you have given, define the term “saturated solution.” In what sense is the term “equilibrium” relevant to what is happening at saturation?

- (d) It is well known that air dissolves in water. By means of sketches and verbal description similar to those you used in parts (a)-(c), describe what happens at the molecular level when an uncovered glass of water, initially free of dissolved air, is put out into the room. At what point does the amount of air dissolved in the water cease increasing?

Note to the student: In the following multiple-choice questions, circle the letters designating those statements you see to be correct. *Any number* of statements may be correct, *not* just one. You must consider each statement on its merits.

13.11 On the scale of relative atomic and molecular masses, nitrogen molecules (N_2) have a value of 28 while chlorine molecules (Cl_2) have a value of 71. With the two gases at the same temperature, the root mean square (rms) value of the velocity of the nitrogen molecules

- (a) is the same as the rms velocity of the chlorine molecules.
- (b) is smaller than the rms velocity of the chlorine molecules by the factor 0.39.
- (c) is larger than the rms velocity of the chlorine molecules by the factor 2.5.
- (d) is larger than the rms velocity of the chlorine molecules by the factor 1.6.
- (e) None of the above.

13.12 If the velocity of every molecule in a fixed volume of gas were doubled,

- (a) neither the temperature nor the pressure would be altered because the volume is kept fixed.
- (b) both the temperature and the pressure of the gas would be doubled.
- (c) the temperature would be quadrupled and the pressure would be doubled.
- (d) both the temperature and the pressure would be quadrupled.
- (e) the specific heat of the gas would be doubled.
- (f) the average momentum of the molecules would be increased.
- (g) few containers would be strong enough to hold the gas.
- (h) None of the above.

Chapter 14

Modern Physics

14.1 The quantity 96,500 C emerges in measurements of amounts of material liberated in electrolysis experiments.

- (a) In your own words, describe and illustrate the meaning of this quantity.
- (b) Once we have accepted the premise that matter is constituted of discrete particles (atoms and molecules), what do the electrolysis measurements imply about how electrical charge is probably parceled out in the structure of matter? Is it likely to be continuous or discrete? Explain your inference carefully. (The inference you have articulated was made by Helmholtz and other scientists around the middle of the 19th century as the atomic-molecular model came to be accepted—a half century before the experiments of Thomson and Millikan.)

14.2 Preceding Roentgen's discovery of X-rays in 1895, the process of electrolysis had been visualized as involving the migration, toward cathode and anode, respectively, of positively and negatively charged atoms (called “ions”) of the elements forming the *compound* being decomposed in the electrolysis.

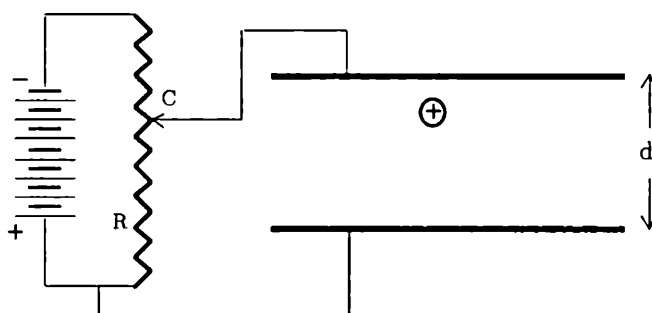
- (a) Explain in your own words how the observations made in electrolysis experiments lead to and support the inference concerning ions. What does the notion (or “model”) of ions suggest concerning how atoms might be held together in forming a molecule of a compound that makes a conducting solution on dissolving in water?
- (b) With the discovery of X-rays, it was soon found that all gases could be made conducting by irradiating them with X-rays. Oppositely charged entities migrated to cathode and anode, even in the case of gaseous *elements* such as hydrogen, helium, neon, oxygen, and nitrogen. Thus the formation of “ions” was not limited to compounds as in electrolysis; elements also formed electrically charged particles. What, if anything, does this discovery add to our view of the electrical structure of matter beyond the inferences already available from the electrolysis experiments?

14.3 Demonstration experiments are frequently performed with Crookes's tubes to illustrate properties of the cathode beam. If you have seen such demonstrations,

consider the following tubes: (1) the tube containing a Maltese cross that can be swung up to intercept the cathode beam or swung down out of the way of the beam; (2) the tube with a small paddle wheel rolling on guides; (3) the tube that exhibits the effect of a magnet on the cathode beam.

- In each of the three demonstrations, describe what was actually *observed* to happen.
- In each of the three demonstrations, describe the *inferences* to be drawn about the properties of the cathode beam.

14.4 A droplet of oil or a small plastic bead, either one having a known density ρ (mass per unit volume) and known radius r , is injected into the space between capacitor plates as shown. The distance between the plates is denoted by d . After the air between the plates has been irradiated with X-rays, the bead becomes electrically charged (either positively or negatively).



- How do you account for the bead becoming electrically charged? (It is a *fact* that it does.) Describe the processes you visualize taking place within the system, starting with the irradiation of the air and ending with excess charge on the bead.
- The potential difference ΔV between the capacitor plates can be varied by moving the contact C up or down along the resistor R, which is connected across the battery as shown in the diagram. Explain in your own words how this part of the system works: Why does the potential difference across the capacitor change as C is moved? Does the potential difference increase or decrease as C is moved up? What is the direction of the electrical field between the plates? Does the electrical field strength become larger or smaller in magnitude as contact C is moved up along R? Explain your reasoning.
- In the diagram, what would be the electrical effect on the charged particle if the contact C were located at the lowest point of the resistor R? Explain your reasoning. With the battery connected as shown, would it be possible to “balance” a negatively charged particle—i.e., keep it from falling under the influence of gravity? Why or why not? How would you connect the battery in order to balance a negatively charged particle?
- Draw a force diagram for the charged particle in the presence of a non zero electrical field and describe each force in words.

- (e) Explaining your reasoning, obtain an algebraic expression for the *weight* of the bead or droplet in terms of the known quantities ρ , r , and g .
- (f) Suppose that it turns out to be possible to “balance” the particle at some measured value of potential difference ΔV_o , i.e., the particle neither falls nor rises at the given electrical field strength. Show that the charge q present on the particle under these conditions can be calculated from the relation

$$q = \frac{4}{3}\pi r^3 \frac{\rho g d}{\Delta V_o}$$

14.5 Millikan, in his measurements of the amounts of electrical charge picked up by oil droplets in air that had been irradiated with X-rays, showed that the charge came in discrete “chunks” or “quanta” and that the size of the smallest chunk observed was 1.60×10^{-19} C.

Thomson had made measurements on positive ions formed in gases through which a cathode beam had been passed and reported that he had been able to observe both singly and doubly ionized species of every gas except atomic hydrogen. With atomic hydrogen he was able to observe only singly ionized atoms, i.e., ions carrying only one quantum of charge. It was also known in chemistry that hydrogen was never found to form a molecule in which two or more atoms of another element combined with only one atom of hydrogen; i.e., hydrogen was to be found only in combinations such as HX or H_nX but never as HX_n (where n denotes an integer number).

- (a) Given the observed facts stated above and the additional fact that 1.01 g of hydrogen is liberated by passage of 96,500 C in electrolysis, calculate the number of *atoms* of hydrogen one would expect to have present in 1.01 g of the gas. Explain your reasoning, indicating the role played by each of the observations.
- (b) Chemists have established that on a scale in which the commonly occurring carbon atom (C) has a relative mass of 12.0, hydrogen atoms (H) have a relative mass of 1.01, sodium atoms (Na) have a relative mass of 23.0, and chlorine atoms (Cl) have a relative mass of 35.5. Suppose we proceed to weigh out 1.01 g of H, 12.0 g of C, 23.0 g of Na, and 35.5 g of Cl. Explain what these very different masses of different materials must have in common. What significance do you now ascribe to the number calculated in part (a)?
- (c) How many atoms altogether (Na and Cl combined) must be present in 58.5 g of common table salt, which has the molecular formula NaCl? How many *molecules* of the combined form NaCl must be present in the 58.5 g? Explain your reasoning in each instance.
- (d) We know that ordinary water has the molecular formula H_2O and that the relative mass of the oxygen (O) atom on the scale cited above is 16.0. How many atoms altogether must there be in 18.02 g of water? How many *molecules* of the form H_2O ? Explain your reasoning.

14.6 The density of solid crystalline sodium chloride (NaCl) is easily measured and is known to be 2.16g/cm^3 .

- (a) Explaining your reasoning, calculate the volume occupied by 58.5 g of NaCl.

- (b) Given the total number of atoms of Na and Cl present in the 58.5 g (using result obtained in Question 14.5), calculate the average spacing between centers of Na and Cl atoms in the solid material. Explain your reasoning with the help of a sketch of the geometry under consideration.
- (c) The result you have obtained in part (b) yields a value for the approximate size (radius or diameter) of Na and Cl atoms. Explain why this is the case, being careful to indicate what role the observed fact that solids are very incompressible plays in this reasoning. Is the number you have obtained to be interpreted as a lower bound, an intermediate value, or an upper bound on the atomic sizes? Explain your reasoning.
- (d) Using the well-known numerical value for the density of liquid water under ordinary conditions, calculate an approximate value for the radius or diameter of the water molecule (H_2O). Use appropriate sketches and explain all aspects of your reasoning, including whether you have obtained an upper or lower bound. Compare the values you have obtained in parts (c) and (d) and comment on the results.

14.7 Consider the following statement: “If, in the Thomson experiment, the two deflecting capacitor plates are brought closer together without changing the potential difference between them and without changing any of the properties of the cathode beam, the deflection of the spot on the screen would be observed to increase.” Is this statement true or false? Explain your reasoning.

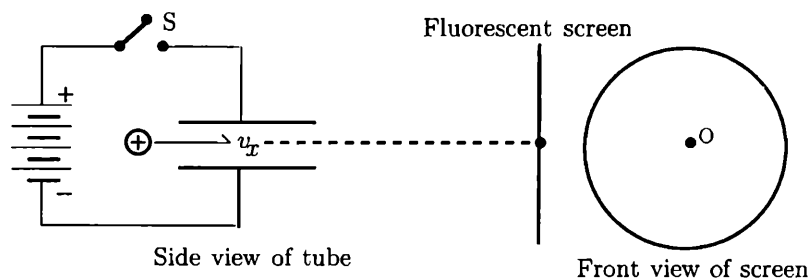
14.8 Consider the following statement: “Gravitational effect (curving toward the ground) of the beam in a cathode ray tube is not observed because the electrons in the beam have so small a mass that the gravitational effect is unobservably small.” Is this statement true or false? Explain your reasoning. If you consider the statement false, how do you account for the fact that the gravitational effect is indeed unobservably small?

14.9 Consider the following statement: “As we decrease the pressure of gas in a discharge tube, we see a larger number of discrete lines in the emission spectrum because, at the lower gas density, the mean free path of the photons increases and more of them are able to get to the walls of the tube and escape from the gas.” Is this statement true or false? Explain your reasoning. If you consider it false, how do you account for the fact that more discrete lines do indeed become visible?

14.10 A beam of positively charged particles passes between capacitor plates in a highly evacuated tube and strikes a fluorescent screen at the end of the tube as shown in the following diagram. When the capacitor plates are uncharged (switch S open), the beam makes a bright spot in the center of the screen at O. The particles in the beam all have the same charge and the same horizontal velocity v_x , but there are two populations of particles with two different masses m_A and m_B , with $m_B > m_A$.

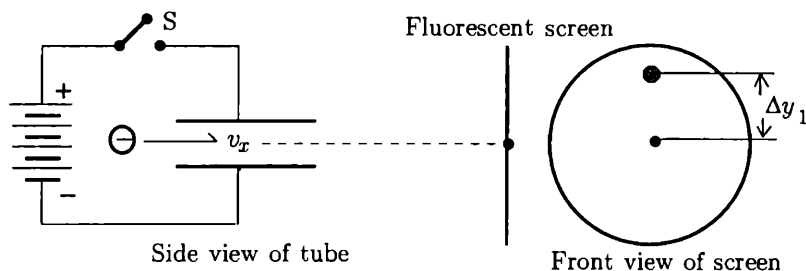
- (a) Sketch on the front view of the screen what you expect to see when the switch S is closed and the capacitor plates become charged. (The potential difference supplied by the battery is sufficient to deflect the beam but is not sufficient to drive the beam off the screen.) Be sure to mark the locations of m_A and m_B . Explain your reasoning.

- (b) Suppose the beam were not homogeneous in velocity (i.e., both populations of particles have velocities ranging from some minimum value $v_{x \min}$ to some maximum value $v_{x \max}$.) Sketch what you would expect to see on the screen under these circumstances with switch S closed. Explain your reasoning.



14.11 Consider an experiment like the Thomson experiment in which a beam of charged particles is passed between capacitor plates and strikes a fluorescent screen at the end of a tube, as shown in the following diagram. A magnetic field can be superposed in a direction perpendicular to the electrical field, i.e., into or out of the plane of the paper.

Suppose a beam of negatively charged particles is homogeneous in charge q and mass m but contains two distinct groups, one, consisting of half the particles, having initial horizontal velocity v_{1x} , and the other group having initial velocity v_{2x} , where v_{2x} is just twice as large as v_{1x} .



The circle at the right shows the screen. When electric and magnetic fields are zero and the beam is undeflected, the beam strikes the spot at the center. The spot a distance Δy_1 above the center is where the v_{1x} particles strike the screen when the switch S is closed and the beam is deflected by the E-field alone.

- (a) Add to the picture whatever is necessary to show what the screen will look like when both groups of particles are present. That is, will there be a second spot or will there be a smearing out of the screen pattern? If there is a second spot, put it on the screen, positioning it correctly in relation to the v_{1x} spot. (The position you show must be correct relative to the scale defined by the length Δy_1 on the diagram, and some ratio reasoning may therefore be required.) Explain your reasoning.
- (b) In the manner exploited by Thomson, a magnetic field is now introduced in a direction perpendicular to the plane of the paper, and its strength is adjusted

(by varying current in the coils producing it) until the v_{1x} particles are returned to the center of the screen. Show what the screen will now look like. Where, that is, will the remaining particles strike the screen? Explain your reasoning.

14.12 In an experiment measuring the photoelectric properties of a certain metal, it is found that the threshold for photoemission is at a certain frequency ν_0 of violet light.

- (a) Suppose the incident light is now changed to a frequency in the ultraviolet region without change in the intensity. What, if anything, will happen to the observed stopping potential and the photocurrent? Explain your reasoning.
- (b) Suppose the incident light is now changed to a frequency in the blue region without change in the intensity. What, if anything, will happen to the observed stopping potential and the photocurrent? Explain your reasoning.
- (c) Suppose the frequency of the incident light is increased to a value $1.43\nu_0$. What will be the maximum kinetic energy of emitted electrons? Explain your reasoning.

14.13 The radius of an atom is of the order of 20,000 times the radius of an atomic nucleus. Let us take the order of magnitude of the density of liquids and solids to be of the order of 2 g/cm^3 . Suppose that atoms were stripped of their cloud of electrons, and the bare nuclei were packed as closely as atoms are in liquids and solids.

Using ratio reasoning, calculate the order of magnitude of the density of matter consisting of closely packed atomic nuclei. (To visualize the significance of your result, recalculate it in units that might be more directly familiar. Try tons per cubic inch, for example.) There is reason to believe that certain stars consist of matter approaching, and even exceeding, this fantastic density.

14.14 After publication of Einstein's suggestion of the photon model, various individuals tried to preserve the older wave picture by proposing modifications of classical theory. Among these efforts was one by J. J. Thomson [*Proc. Camb. Phil. Soc.*, XIV, 417 (1907)]. Thomson suggested that electromagnetic energy might be unevenly distributed over the wave front, with regions of maximum energy relatively widely separated by areas of low or zero disturbance. This hypothesis led to the suggestion that at extremely low light intensities, when only a few maximum energy regions would be present on any given wave front, ordinary diffraction patterns formed by slits or shadows of small obstacles might be modified in some observable way, perhaps by fuzzing out of the pattern.

An experiment to test this hypothesis was carried out by G. I. Taylor [*Proc. Camb. Phil. Soc.*, XVI, 114 (1909)]. Taylor reports that:

Photographs were taken of the shadow of a needle, the source of light being a narrow slit placed in front of a gas flame. The intensity of the light was reduced by means of smoked glass screens. . . . The longest time [of exposure with very weak light] was 2000 hours or about 3 months. [Legend has it that Taylor, being an avid sailor, went off sailing during this period.] In no case was there any diminution in the sharpness of the pattern. . . . The amount of energy falling on one square cm of the

plate is 5×10^{-6} erg/sec, and the amount of energy per cubic cm of this radiation is 1.6×10^{-16} erg.

- (a) Show that from the point of view of the photon model, the given energy flux of 5×10^{-6} erg/(s)(cm²) corresponds to about 106 photons of visible light per second per square centimeter (taking the average energy per photon to be about 2 or 3 eV).
- (b) Show that this flux of photons implies that the average distance of separation between individual photons must have been of the order of 300 m.

With an apparatus of the order of 1 m in length, it is extremely unlikely that more than one photon would have been present in the system at the same time! With modern counting equipment and photomultiplier tube detectors, Taylor's experiment is readily repeated without waiting 3 months for photographic exposure. All such experiments confirm Taylor's original results. For example, a two-slit interference pattern formed by an intense beam of light is identical with one formed by light so weak that only one photon is likely to be passing through the slit system at any given instant. Closing either one of the two slits eliminates the interference pattern.

- (c) What do you infer from the fact that an interference pattern is still formed under conditions of extremely weak illumination? How can two slits be effective in producing an interference pattern when only one photon arrives at a time? [Do not expect to give or receive a simple, pat answer to this question; the problem being posed lies at the heart of modern quantum physics, but there is every reason for you to begin to think, and wonder, and speculate about it.]

Note to the student: In the following multiple-choice questions, circle the letters marking those statements that are true or correct. *Any number* of statements may be correct in a given question, *not* necessarily just one. It is necessary to examine each statement on its merits.

14.15 The measurements of alpha particle scattering made by Geiger and Marsden and interpreted by Rutherford showed for the first time that

- (a) protons are more massive than electrons.
- (b) atoms contain a very dense concentration of positive charge.
- (c) neutrons must be present in atomic nuclei.
- (d) alpha particles must be positively charged.
- (e) the region of dense concentration of positive charge must be very much smaller than the size of the atom.
- (f) oscillating electrically charged particles must radiate electromagnetic waves.
- (g) electrons circulate around the nucleus in circular or elliptical orbits.
- (h) alpha particles are scattered by target atoms through large as well as small angles.
- (i) None of the above.

14.16 [In this question you must be careful to discriminate between what is observed and what is inferred from observations.] In the photoelectric effect, light incident on a metal surface causes electrons to be ejected from within the metal. If the intensity of an incident beam of monochromatic light is decreased, it is an OBSERVED FACT that:

- (a) The rate of ejection of electrons decreases.
- (b) The photoelectric current ceases below a certain threshold of minimum intensity of the incident light.
- (c) The maximum kinetic energy, possessed by the ejected electrons at the point at which they leave the metal surface, decreases.
- (d) The observed photoelectric current decreases.
- (e) The potential difference that just brings the photoelectric current to zero remains unchanged.
- (f) Electrons are bound within the metal by a certain minimum amount of energy referred to as the “work function.”
- (g) None of the above.

14.17 Millikan, in his famous oil drop experiment,

- (a) measured the charge carried by the cathode ray particles (electrons) detected in discharge tubes such as those used in the Thomson experiment.
- (b) demonstrated that both positive and negative electrical charges are quantized, i.e., come in discrete packages of identical finite size.
- (c) made it possible to calculate a reasonably precise value of Avogadro's number.
- (d) demonstrated that atoms have very tiny positively charged nuclei with electrons located relatively far from the nucleus.
- (e) took into account the frictional effect between the droplet and the surrounding air in those observations in which terminal velocities of the droplets were being observed.
- (f) neglected the effect of gravity compared with the electrical force on the charged droplet.
- (g) had to find the smallest common multiple among the charge quantities he measured because the droplets, on different occasions, carried varying numbers of “packages” of charge.
- (h) None of the above.

14.18 In the simplest Bohr model of the hydrogen atom, with the electron visualized as executing purely circular orbits around the proton:

- (a) The resulting formula for the allowed orbits shows that the radii of these orbits increase in equal steps of length with increasing value of the number n .
- (b) Neglect of the gravitational force between the electron and proton is one of the basic faults of the model.
- (c) A “stationary state” is one in which the electron has zero velocity.
- (d) The electron orbit should be surrounded by a magnetic field similar to that surrounding a current-carrying loop of wire, and each hydrogen atom should therefore behave like a minute magnet with north and south poles.
- (e) In the expression $E(r) = -ke^2/2r$ for the total energy associated with an orbit of radius r , the 2 in the denominator arises in the calculation of the potential energy associated with the Coulomb interaction in the electron-proton system.
- (f) A higher energy photon is released when the electron jumps from orbit $n = 2$ to orbit $n = 1$ than is released when the electron jumps from orbit $n = 100$ to orbit $n = 2$.
- (g) The kinetic energy of the electron is larger in orbit $n = 4$ than in orbit $n = 2$.
- (h) The orbital kinetic energy of the electron approaches zero as the radius of the orbit increases without limit.
- (i) The negative sign associated with the total energy of the system in any allowed orbit stems from the fact that the potential energy decreases more than the kinetic energy increases as the radius decreases from very large values.
- (j) In the “ground state” of the atom, the radius of the electron orbit is essentially zero.

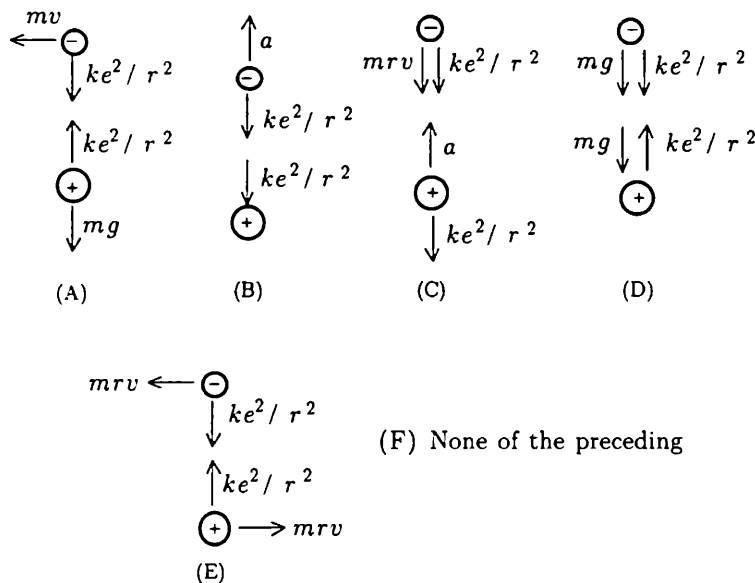
14.19 According to the Bohr model of the hydrogen atom:

- (a) A photon of energy $h\nu = 2\pi^2m(ke^2)^2/h^2$ would be capable of ionizing a hydrogen atom.
- (b) The reddest line in the Balmer spectrum involves an electron transition from orbit 3 to orbit 2.
- (c) Transitions are possible between any two allowed orbits.
- (d) If an incident photon, absorbed by an electron in the ground state, has just the right energy to kick the electron from orbit 1 to orbit 5, as many as four different photons might be emitted as the electron cascades back to the ground state.
- (e) The electron would have nonzero angular momentum in all allowed states including the ground state.
- (f) None of the above.

14.20 How are some of the X-rays produced in a cathode ray tube in which the electron beam strikes a metallic target?

- (a) The electrons, when accelerated through a large potential difference, are converted into X-rays on striking the target.
- (b) Photons are accelerated by the high potential difference until their energy lies in the X-ray region.
- (c) The electrons emit X-rays as they are accelerated initially.
- (d) The electrons emit electromagnetic radiation when they are slowed down on striking the target.
- (e) X-ray photons, present in the target, are released when electrons strike the target.
- (f) X-ray photons are liberated from the target by the photoelectric effect.
- (g) Electrons annihilate positrons in the target, producing electromagnetic radiation.
- (h) None of the above.

14.21 In the Bohr model of the hydrogen atom, an appropriate force diagram (or diagrams) for the two interacting particles would be:



14.22 A beam of X-rays passes near a charged electroscope. The leaves of the electroscope will collapse if the electroscope is

- (a) confined in a highly evacuated chamber.
- (b) surrounded by a gas.
- (c) charged positively, but not if it is charged negatively.

- (d) charged negatively, but not if it is charged positively.
- (e) not subject to a magnetic field.
- (f) None of the above.

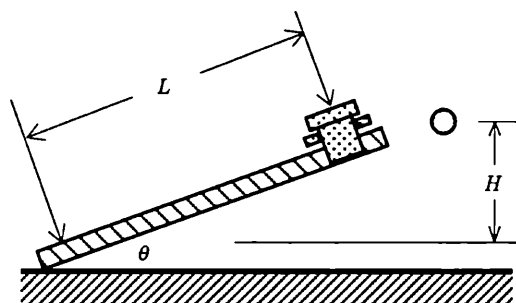
14.23 It is an observed fact that the emission spectrum of an atom (e.g. sodium) has many more bright lines than the absorption spectrum has dark lines. This difference between emission and absorption spectra is explained by the fact that:

- (a) Fewer atoms are absorbing light when absorption spectra are formed than are emitting light when emission spectra are formed.
- (b) In absorption, electrons are not elevated to energy levels as high as those from which they start in the case of emission.
- (c) In absorption electrons are all elevated to higher energy levels from the ground state while in emission they cascade back to the ground state through intermediate states.
- (d) The ionization potential of the atoms is a variable quantity.
- (e) The magnetic moments of the atoms are different in different energy levels.
- (f) None of the above.

Chapter 15

Mixed Areas of Subject Matter

15.1 The diagram shows a system in which a glider with mass m_G , starting from rest, slides down an air track inclined at an angle θ to the horizontal. It descends along a length L of the track while dropping through a vertical height H . Simultaneously, a ball with mass m_B also starting from rest, drops vertically through the same height H . We shall compare these two motions in two different ways and analyze their similarities and differences. (Frictional effects are to be taken as negligible in the analysis, which is to be carried out entirely algebraically.)

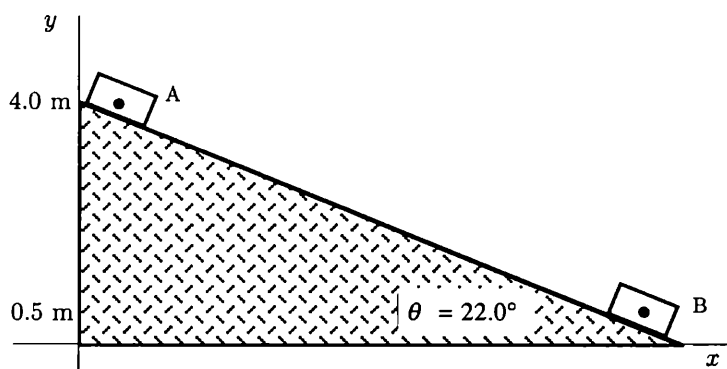


- (a) What is the linear acceleration of the glider? How does it compare in magnitude with that of the ball? Explain your reasoning.
- (b) Use the kinematic equations to obtain expressions for the velocities v_G and v_B of the glider and the ball after each has dropped through the height H , as well as expressions for the time intervals Δt_G and Δt_B for the descent of each object to this final level.
- (c) Use a conservation of energy argument to obtain expressions for v_G and v_B .
- (d) Analyze and interpret the results you have obtained in parts (b) and (c). How do the two velocities compare? Which approach is the more powerful for obtaining the velocity information? How do the two time intervals compare? That is, which is longer? How do you explain the difference? Analyze the time difference algebraically: How does it vary as the angle of inclination of the track is changed without changing H ? Show that

$$\Delta t_G - \Delta t_B = \sqrt{\frac{2H}{g}} \cot \theta$$

- (e) Could you have obtained the time difference information directly from the energy argument alone? Why or why not?

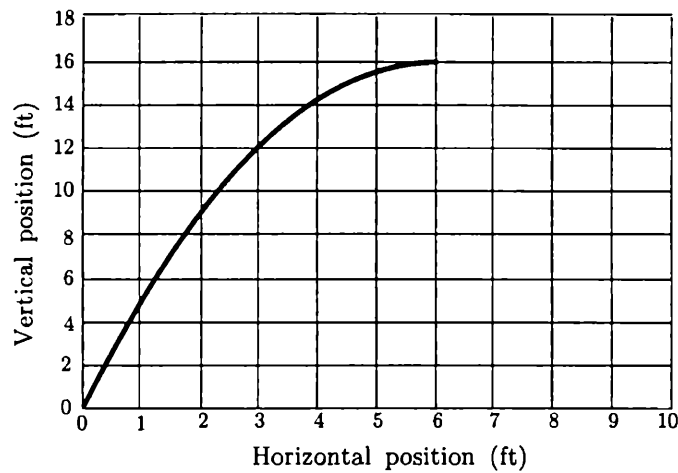
15.2 A plane is inclined at an angle of 22.0° to the horizontal as shown. A block with a mass of 12.0 kg, sliding down the plane under the influence of friction, passes position A, with its center of mass at a height of 4.0 m above the ground, and with a velocity of 3.52 m/s. It slows down under the influence of the substantial frictional force, coming to a stop at position B, with its center of mass 0.5 m above the ground.



- (a) Draw force diagrams of (1) the block at some intermediate position between A and B and (2) of the region of the plane in contact with the block at that same position. Describe each force in words.
- (b) Explaining your reasoning, calculate the increase in thermal internal energy (between the initial and final conditions) of the system consisting of the block and the plane, assuming that there is negligible heat transfer to the surrounding air as the temperatures of the block and plane increase.
- (c) Between the initial and final conditions, what is the total change in internal energy of the entire system consisting of the earth, the block, and the plane? Explain your reasoning.

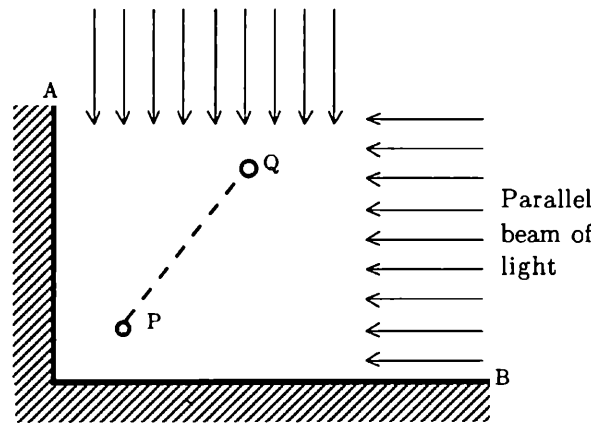
15.3 A ball is projected in a trajectory such as that sketched. Exactly at the top of its flight it undergoes a perfectly *inelastic*, head-on collision with an identical ball that is suspended on a very weak thread, which breaks on the collision.

- (a) Sketch on the diagram the continuing trajectory of the two-ball combination after the top of the flight to the point at which it lands on the ground. Explain how you drew the trajectory.
- (b) Defining the system clearly and carefully, describe all the momentum and energy changes that take place as the first ball is on the way up, as the two balls collide, and as the two balls continue to the ground.



Note to the instructor: Question 15.4 is intended to help students understand the concept of “components of a displacement vector” by connecting this idea to the displacement of the shadow of an object along a wall.

15.4 Consider the situation illustrated in the following diagram: Two walls, A and B, are shown perpendicular to each other. A small ball is located initially at point P. Two fairly distant search lights (one directing a parallel beam directly at wall A, and the other a parallel beam directly at wall B) cast sharp shadows of the ball on each wall.



- (a) Denoting the shadow points by the symbols P_A and P_B , respectively, mark the location of the shadow of the ball on each wall when the ball is at point P.

The ball is now moved from P to Q, a distance of 3.76 m, along the dashed line. This line makes an angle of 52.5° with wall B and lies in the plane that is represented in the diagram.

- (b) Denoting the shadow points by the symbols Q_A and Q_B , respectively, mark the location of the shadow of the ball on each wall.

- (c) Calculate the distances P_AQ_A and P_BQ_B , i.e., the distances between the successive shadows on each wall. Explain your reasoning.
- (d) Return to the concepts of “displacement vector” and “rectangular componer of a displacement vector” that you studied in mechanics. Explain the connections between the calculations you made in part (c) and these vector concep Explain why components of velocity vectors can be similarly calculated.

Note to the instructor: In general, students have not been led to look at tabl of data to discern significant relations, insights, or generalizations that are frequent contained in the data. They tend to view tables as useful only for finding numbe to put into end-of-chapter problems. Following is an example of what might be do with a fairly mundane table, that of densities of various materials.

15.5 This table gives the densities of various materials at room temperature (20 ° and atmospheric pressure.

Table of Densities (g/cm ³)					
Solids		Liquids		Gases	
Material	Density	Material	Density	Material	Density
Aluminum	2.70	Carbon tet- rachloride	1.59	Air	1.20×10^{-3}
Brass	8.5	Ethyl alcohol	0.791	Ammonia	0.77×10^{-3}
Copper	8.96	Gasoline	0.66-0.69	Carbon monoxide	1.84×10^{-3}
Gold	19.3	Mercury	13.6	Carbon dioxide	1.84×10^{-3}
Iron	7.87	Methyl alcohol	0.810	Chlorine	3.00×10^{-3}
Lead	11.4	Milk	1.03	Helium	0.165×10^{-3}
Platinum	21.4	Oils	0.92-0.97	Hydrogen	0.084×10^{-3}
Silver	10.5	Seawater	1.025	Methane	0.67×10^{-3}
Sodium	0.97	Sulfuric acid	1.847	Oxygen	1.33×10^{-3}
Tin	5.75	Water	1.00	Nitrogen	1.16×10^{-3}
Uranium	19.0	Liquid air (at -193°C)	0.87	Sulfur dioxide	2.73×10^{-3}
Zinc	7.13			Air at 10 km*	0.37×10^{-3}
Ice	0.917			Air at 20 km*	0.09×10^{-3}
Glass	2.4-2.8				
Plastics	0.90-1.1				
Salt	2.18				
Sugar	1.59				
Stone	2.4-3.1				
Wood					
Balsa	0.11-0.14				
Oak	0.60-0.90				

*in situ values, not at
20 °C and 1 atm.

- (a) Examine the table from the standpoint of comparing the general categories of solids, liquids, and gases: Is there a general trend of increase or decrease in density as one shifts from one category to another? Which way? To what extent, if any, do the categories overlap, or not overlap? What, if anything, seems special about gases? What is the *order of magnitude* (round number) of the ratio of the density of gases to that of liquids?
- (b) What are the most dense materials listed? The least dense? Are you surprised by the position of lead in the sequence? Suppose you put a piece of lead in a beaker of mercury. How would the lead behave? How many of the listed materials would sink in mercury? Are there any metals that would float on water (if prevented from reacting explosively with it)?
- (c) Suppose you wished to concoct a liquid that is not water but has a density very nearly equal to that of water. On the basis of the table, what liquids might you try to mix together to achieve this? If you were actually to undertake such a task, what had you better find out about various properties and behaviors of the liquids you wish to mix?
- (d) Air is actually a mixture of the gases oxygen and nitrogen (roughly 20% oxygen). Suppose the gases separated in the atmosphere instead of remaining mixed. Which gas would end up “floating” on top?
- (e) Scientists of the late 18th and early 19th centuries found it hard to see why the gases did not separate, with the nitrogen floating on top. The fact that the gases remained mixed puzzled them. From your present vantage point, how would you explain the situation to one of those scientists and make your explanation convincing?

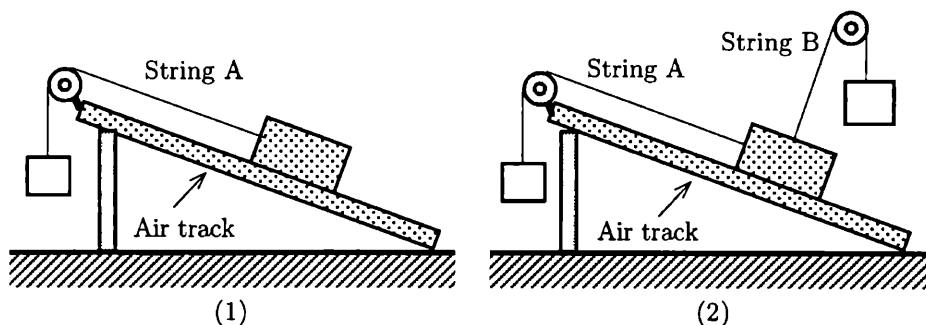
Note that the density of gases is very, very much less than that of liquids or solids. Note also that gases are highly compressible, very much more so than liquids or solids. [You are aware of this although you may never have articulated it explicitly. “Highly compressible” means that it is easy to decrease the volume of a gas very substantially by squeezing it (as in a bicycle pump), but it is extremely difficult to change the volume of liquids or solids by squeezing. The volume of the latter can be decreased somewhat, but enormous pressures are required to produce significant changes. Liquids and solids are therefore described as “relatively incompressible.”]

- (f) If we accept the view that material substances consist of discrete particles (atoms and molecules), what do the observed densities and facts about compressibility suggest about the relative spacing of particles in gases on the one hand and liquids and solids on the other? Explain your reasoning. In the light of the round number density ratio you noted in part (a), what is the order of magnitude of the ratio of average spacing between particles in gases at room temperature and atmospheric pressure on the one hand and the average spacing in liquids or solids on the other? Explain your reasoning.

Note to the instructor: Question 15.6 connects thinking in electrical and mechanical contexts.

15.6 The direction of the electrical field everywhere along the surface of a charged metallic conductor must be normal to the surface after the charge has settled down into an equilibrium distribution.

- (a) Why is this the case? What would happen to any excess charge in a region of the surface where the field direction was not normal to the surface? (Draw simple diagrams to illustrate your argument.)
- (b) Must the electrical field direction be everywhere normal to the surface of a charged nonconductor? Why or why not?
- (c) Compare your argument about the electrical situation in the charged metallic conductor with the mechanical situation shown in figure (1) in which a block is held at rest on a sloping air track (negligible friction) by means of the balancing force exerted by string A parallel to the track: What must be the direction of the force exerted by the air track on the block under this equilibrium condition? Why is this force given the name “normal force”? What would be the effect on the block if this force were not acting in a direction perpendicular to the track but had a component one way or the other along the track?



Suppose the effect of the track in figure (1) is replaced by the pull of string B as shown in figure (2), and the block is just barely lifted off the track.

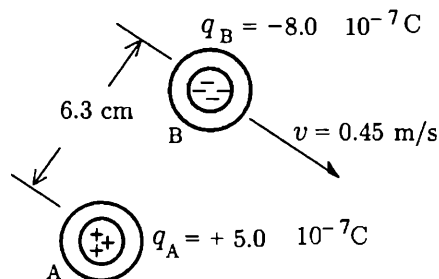
- (d) What is the equilibrium orientation of string B? Suppose you displace the block slightly up or down the plane from the equilibrium position so that string B is no longer normal to the track. What will happen to the block after you let it go from the displaced position?
- (e) Explain the analogy between this mechanical situation and that of a charged region in the metallic conductor where the electrical field is not initially normal to the surface.

Note to the instructor: Question 15.7 has several purposes. One is to spiral back to the centripetal force ideas as soon as Coulomb's law is available. Another is to confront students with a situation in which the applied force may be either larger or smaller than that necessary to impart the indicated centripetal acceleration, and the motion may not then be circular. (Students rarely encounter such situations, and they develop the habit of substituting indiscriminately into the formula applicable to circular motion as though it were universally valid.) The third purpose is to present a situation in which the student is not explicitly told what to calculate and must make the decision independently. This is a very rarely available opportunity.

15.7 We are looking down on two small frictionless pucks A and B located on a level air table. Firmly fastened to each puck is a small uniformly charged sphere, as

indicated. Puck A is firmly fastened to the table and does not move. Its sphere carries a charge of $+5.0 \times 10^{-7} \text{ C}$. The sphere on puck B carries a charge of $-8.0 \times 10^{-7} \text{ C}$, and the mass of B (with its charged sphere) is 125 g. At a given instant, puck B has an instantaneous velocity of 0.45 m/s in a direction perpendicular to the radial line between A and B. The radial separation of the pucks at this instant is 6.3 cm.

Investigate what will happen after the condition at the instant shown: Will puck B follow a circular path around A? If not, will it tend to increase or decrease its radial distance from A? Note that you are not being told exactly what to calculate; you must make the decision yourself. That is one of the main points of the problem. Show all your numerical calculations, explaining why you make each one and explaining the inferences to be drawn from your calculations. Handle units carefully. Be sure to draw force diagrams for both pucks A and B.



15.8 A heating element with a resistance of 25Ω is designed to be connected to a 120 V source of electric current.

- (a) Explaining your reasoning by giving relevant definitions, calculate the wattage rating of the heating element.

The heating element is now immersed in 1000 g of water in a well-insulated brass calorimeter at an initial temperature of 20°C and is allowed to run for 15.0 min, after which it is removed from the water. The brass container has a mass of 250 g and a heat capacity of $0.10 \text{ cal}/(\text{g})(^\circ\text{C})$.

- (b) Explaining your reasoning as you go along, predict the final condition in the calorimeter. That is, what will be the final temperature? Will any of the water have been boiled away? If so, how much?

15.9 In our investigation of the magnitude of atomic and molecular dimensions, we found that these were of the order of 2 or 3 angstroms. In measurements of the wavelength of visible light, we have found such wavelengths to be of the order of 5000 angstroms. Recall what happens to a wave train that passes by an object much smaller than its wavelength.

In the light of the insights and experiences listed above, discuss the possibility of our actually seeing an atom or molecule by illuminating it with *visible* light. Appeal, for example, to what you might have observed when water waves or ripples are incident on large or small obstacles. Explain your reasoning carefully.

15.10 Through observing objects or systems “doing things” to each other, we become aware of interactions in which accelerations can be imparted. Through Newton’s first law, we associate such accelerations with the action of forces, whether or not we observe direct physical contact between the interacting objects. We have, so far, in macroscopic phenomena, separated observed interactions not involving direct physical contact into three classes having the names “gravitational,” “electric,” and “magnetic.” At this point let us review these effects and see them in perspective.

By listing (1) conditions under which each of these interactions is observed, (2) effects and changes we can produce through our own manipulation of the interacting objects, and (3) specific differences we discern among the interactions, show why we are justified in distinguishing three different classes of phenomena rather than lumping them into one. Be specific in recognizing similarities as well as differences. Your listing should be highly “operational”; in other words, you should describe specific things we can do and the consequences of doing them.

Some hints as to relevant aspects: What specific substances are involved? How do different substances differ in their behavior under relevant treatment? Are relevant properties transferable (or not transferable) from one object to another on contact? What changes can be effected by touching or otherwise manipulating the interacting objects? What evidence is there for presence or absence of “induction” effects? Are there differences between any of the observed induction effects? And so forth.

15.11 A formula, standing by itself, is nothing but a collection of letters or symbols. A formula is given meaning only by the *text* that goes with it. The text contains (1) a description of the meaning of the symbols, (2) the physical situations to which the formula does (and does *not*) apply, and (3) the respective roles of definition, empirical result of experiment, or broader “law of nature” that enter into or are expressed in the formula.

Write a paragraph about each of the following formulas, describing its meaning in terms of the characteristics cited above. Take into account the fact that some formulas are true because all the quantities that enter into them are defined and no knowledge of behavior in nature is therefore necessary to make them true. Such formulas may apply, however, to some natural phenomena.

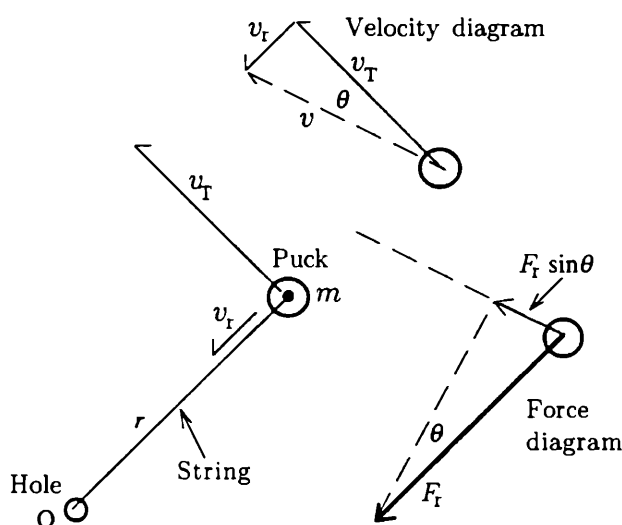
- | | |
|---------------------------------------|-------------------------------------|
| (a) $v = v_o + at$ | (f) $s = s_o + v_o + (1/2)at^2$ |
| (b) $\vec{F}_{\text{net}} = m\vec{a}$ | (g) $F = kx$ |
| (c) $f_{\text{max}} = \mu N$ | (h) $F_{\text{cent}} = mv^2/r$ |
| (d) $F_{\text{elect}} = kq_1q_2/r^2$ | (i) $F_{\text{grav}} = Gm_1m_2/r^2$ |
| (e) $pV = \text{constant}$ | (j) $\Delta V = IR$ |

15.12 Consider the following classes of uniform circular motion: (1) A bob on a string in a circle lying in a horizontal plane or a frictionless puck tied to a fixed point on a level air table and (2) a satellite in circular orbit around the earth or a negatively charged particle in circular orbit around a fixed positively charged particle.

By appealing to appropriate physical laws and the resulting equations, show that in case (1), any angular velocity is possible at a given radius up to the point at which the string breaks, but that in case (2) only one angular velocity is possible at a given radius. Explain this profound difference between the two cases.

15.13 Suppose we have a frictionless puck of mass m attached to a string on a level air table, as shown. The puck revolves in a circle around point O where there is a hole in the table. The string can be pulled down through the hole at O, decreasing the radius r of the circle in which the puck moves. The tangential velocity of the puck at any radius r is denoted by v_T . Suppose we pull down on the string, decreasing r very slowly, and imparting a very small radially inward velocity v_r to the puck in addition to its tangential velocity v_T .

We know that the puck speeds up as we shorten the radius, increasing its tangential and angular velocities in a manner analogous to the increase in angular velocity of a skater pulling in his or her arms. But note that there is something paradoxical about this situation: There is no external torque applied to the system of puck and string since the force is directed along the string and has no component perpendicular to it. How does the puck manage to speed up in the absence of external torque? Is there something wrong with our dynamical theory? In this problem, we shall analyze, in detail, the speeding up of the puck. First, construct for yourself the force and velocity diagrams sketched in the figure.



In the velocity diagram for the puck, the resultant of velocities v_T and v_r is a velocity v in a direction lying at a very small angle θ below the tangential direction. We shall call *theta* the “angle of descent,” and the direction of descent is shown by the dashed line. (The idea is to keep v_r *very* much smaller than v_T ; the scale of the diagram is greatly exaggerated to make the effect of v_r clearly visible.)

The force diagram for the puck is also shown. Note that the centripetal force F_r has a nonzero component in the direction of descent, i.e., along the resultant velocity v . It is this component of the force (in the presence of the small radial velocity v_r) that accelerates the puck. Now we know, qualitatively, how the puck gets to be accelerated in the absence of an external torque, but we do not have a quantitative analysis. What we want is an equation relating v_T and r directly so that we can see what happens to the tangential velocity as we pull the puck in or let it out. Let us analyze the situation algebraically, starting with the acceleration imparted when v_r is not zero. Work out the algebraic details in each of the following steps carefully with your own pencil and paper.

- (a) Let us denote the acceleration in the direction of descent (i.e., the direction of the dashed line) by a_D . Noting how the accelerating force is related to the centripetal force on the force diagram, argue that

$$a_D = \frac{v_T^2}{r} \sin \theta \quad (1)$$

- (b) We can replace $\sin \theta$ by its connection with the velocities v_r and v on the velocity diagram, but we must now be very careful about algebraic signs because v_r can be either positive (outwardly directed) or negative (inwardly directed). From the velocity diagram show that we can write

$$a_D = -\frac{v_T^2}{r} \frac{v_r}{v} \quad (2)$$

where we have introduced the minus sign to keep a_D positive when v_r is inward and negative when v_r is outward.

- (c) If v_r is very small (i.e., if we pull the puck in or let it out very slowly), the resultant velocity v is very nearly equal to v_T . Argue that we can alter Eq. 2 to

$$a_D \cong -\frac{v_T v_r}{r} \quad (3)$$

- (d) Since our objective is to obtain a relation between v_T and v_r , we want to get rid of a_D in terms of these quantities. For this purpose we can make powerful use of the chain rule of differentiation:

$$a_D = \frac{dv}{dt} = \frac{dv}{dr} \frac{dr}{dt} \quad (4)$$

Show that Eqs. 3 and 4, combined with the fact that v is very nearly equal to v_T , yield the simple differential equation:

$$\frac{dv_T}{dt} \cong -\frac{v_T}{r} \quad (5)$$

- (e) Solve the foregoing differential equation (Eq. 5) to show that the quantity rv_T is constant and this, in turn, implies that $r^2\omega$ is constant (where ω is the angular velocity of the puck). In other words

$$r^2\omega = r_o\omega_o^2 \quad (6)$$

where r_o and ω_o denote, respectively, any arbitrary initial starting radius and angular velocity. Interpret Eq. 6: What happens to the angular velocity as the puck is pulled inward? As it is let outward? Is this consistent with the effect observed in the whirling skater?

- (f) Argue that the analysis we have carried out shows that the angular momentum of the system is conserved, as it should be in the absence of external torque, but that the puck is nevertheless accelerated tangentially as it is slowly moved radially.
- (g) Carry out an integration to obtain an expression for the work done on the system in pulling the string inward (or letting it out), and show that this amount of work turns out to be equal (as one might expect) to the change in kinetic energy of the puck. (Be very careful about algebraic signs.)
- (h) Why has our analysis been confined to pulling the puck in or letting it out very slowly? How does the puck behave if we pull it in rapidly? (Don't try to analyze the latter situation algebraically; just try to visualize it physically—or, better yet, try it out experimentally.)

15.14 The following table summarizes typical values of energy associated with various particles, photons, and physical changes.

Comparison of Energies Associated with Various Microscopic Particles and Phenomena			
Type of energy	Substance, particle, or photon	Energy in electron volts (eV)	Wavelength λ of photon of same energy (angstroms)
Average total energy of vibration of an atomic particle along one coordinate axis in a solid material at room temperature	—	0.02 (approx.)	
Average total translational kinetic energy of gas molecule at room temperature	—	0.04 (approx.)	
Average energy released per atom in violent chemical reaction (e.g., the explosion of TNT)	—	0.5 (approx.)	
First ionization potential (energy necessary to separate one electron from an isolated neutral atom)	K Na Ag Pt Au H He	4.3 5.1 7.5 8.9 9.2 13.6 24.5	505
Photoelectric work function (minimum energy necessary to eject electron from from surface of a metal)	K Na Ag Au Pt	3.0 3.0 4.7 4.8 6.3	
Photons of various types of electromagnetic radiation	Far infrared Sodium D lines Ultraviolet X-rays	0.02 2.1 6 Tens of thousands	5890
Radioactive emanations	α particles from U and Ra β particles from many isotopes γ rays from many isoptoes	4-6 MeV 1-2 MeV 0.01-1 MeV	

You can begin to understand and appreciate the pattern and scale of energy transformations in the microscopic domain by viewing and assimilating the various orders of magnitude exhibited in this table. The simple calculations and questions in parts

(a) through (f) will help you begin to familiarize yourself with the table and its significance.

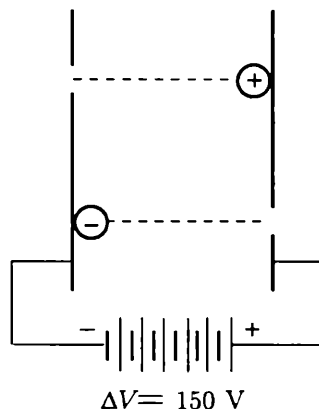
- (a) Fill in the blanks in the right-most column (column 4) by calculating the corresponding value of wavelength λ for each instance in which a value is not already entered.
- (b) We associate the increase in temperature of any material with an increase in the kinetic (or “thermal”) energy of the atoms or molecules of the material. Note the order of magnitude of such energies as indicated in the first two entries of column 3. Note also that your calculations in column 4 show that photons of such energies have wavelengths in the far infrared. Why do you suppose that infrared radiation is frequently referred to as “thermal radiation”?
- (c) The average translational kinetic energy E of gas molecules is given as a function of absolute temperature T by the following equation: $E = 1.29 \times 10^{-4}T$, where E is in electron-volts. At what temperatures would the average kinetic energy of gas molecules be in the range of energies of photons of visible light? (Answer: Temperatures of the order of 15,000 K.) Note that the temperature of the surface of the sun is about 6000 K.
- (d) Note the enormous difference between the order of magnitude of energy per atom in a violent chemical reaction and that associated with radioactive disintegration. What implications do you see in this difference?
- (e) How do you account for the fact that X-rays and radioactive emanations cause ionization of gases while visible light does not? What implications does this have with respect to the effects one might expect X-rays and radioactive emanations to have on complex molecules such as those that, for example, make up biological materials?
- (f) The work function for ejecting electrons from a metal is lower than the ionization energy of an isolated atom of the same substance. What implications do you see in this difference?

You would do well to note and remember the various orders of magnitude illustrated in this question. This will help you think qualitatively and powerfully about a vast range of physical phenomena occurring in the world around us.

15.15 You are informed that in an empty compartment, to which you have access with instruments, there exists either an E-field or a B-field but not both. Describe several experiments or observations you might perform in the empty space to determine which kind of field is present.

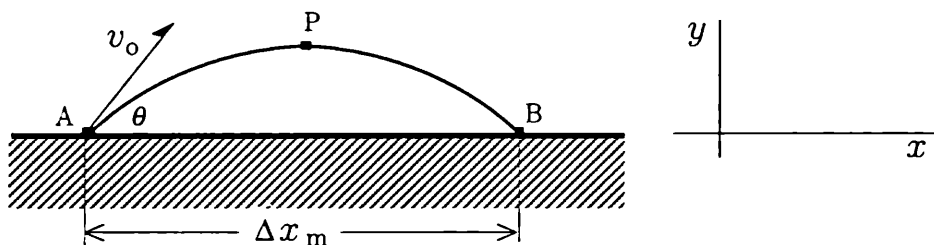
15.16 Consider the following thought experiment: An electron and a proton (hydrogen ion), released from rest at opposite plates of a capacitor in a highly evacuated space, are accelerated across the capacitor and exit through holes in the plates as shown in the following diagram. The potential difference between the plates is 150 V. The mass of a proton is very much larger than the mass of an electron.

- (a) Explaining your reasoning, compare the kinetic energies of the two particles as they exit through their respective openings. ("Comparing" means establishing whether one quantity is greater than, equal to, or smaller than the other.) It is not necessary to calculate numerical values of the kinetic energies.
- (b) Explaining your reasoning, compare the momenta of the two particles as they exit through their respective openings. It is not necessary to calculate numerical values of the momenta.



Note to the student: In Question 15.17 circle the letters marking all those statements that are correct. *Any number* of statements may be correct, and each one must be examined carefully on its merits. Do not simply abandon the question when you have found one correct statement.

15.17 A projectile fired from point A at ground level has initial velocity v_0 inclined at angle θ to the horizontal, as shown in the following diagram. If the air resistance were negligible, the projectile would follow a parabolic trajectory through point P, returning to the ground at point B. The range of the trajectory is denoted by Δx_m . The following statements are to be taken as applying to this idealized situation.



- (a) The magnitude of the instantaneous velocity at point P (top of the flight) is given by $v_0 \cos \theta$.
- (b) If Δt_m denotes the total time of flight of the projectile, the range Δx_m is given by $(v_0 \cos \theta) \Delta t_m$.
- (c) The time Δt_P between the firing of the projectile and its arrival at point P is given by $(v_0 \sin \theta)/g$.
- (d) The acceleration of the projectile has the constant magnitude g throughout the entire history of the flight.
- (e) The magnitude of the velocity of the projectile is larger at point B than it is at point P.

- (f) The magnitude of the velocity of the projectile is the same at point B as it is at point A.
- (g) The momentum of the projectile is not conserved because the projectile does not constitute a closed system.
- (h) The momentum vector of the projectile is rotating in the clockwise direction throughout the flight.
- (i) The momentum vector of the projectile has its smallest magnitude at point P.
- (j) The kinetic energy of the projectile changes during the course of the flight but is the same at point B as it was at point A.
- (k) When the projectile strikes the ground at point B, its kinetic energy is entirely converted into thermal energy within the projectile and within the ground in the vicinity of the impact.
- (l) None of the above.

Chapter 16

Naked Eye Astronomy

16.1 Describe how you would go about establishing vertical direction and horizontal direction, using the simplest possible materials or devices, at any place you happen to be located, especially in the case of a sloping hillside. (You are being asked, in other words, to give simple operational definitions of the terms “vertical” and “horizontal.”)

16.2 Define the terms “local zenith point” and “local celestial meridian” and go on to do the following.

- (a) Describe how you might go about identifying the direction of the local zenith point and mapping out the local celestial meridian at any place you happen to be located. Note that there are two different ways of doing the latter: One is to make use of the North Star, and the other is to make use of the shadow, cast by the sun, of a vertical stick. (It is legitimate to think of carrying out these tasks fairly crudely; high precision is not required.)
- (b) Define the term “local noon” by describing how you would determine the moment of local noon without reference to a clock.
- (c) Account for the fact that the actual moment of local noon varies continuously from east to west over any conventional time zone.

16.3 Give an operational definition of the term “geographic north-south direction” by describing how you would go about establishing this direction at any point at which you happen to be located.

In the light of the universally accepted definition of “geographic north-south direction,” explain in your own words why the magnetic compass does *not* serve to *define* this direction. What *is* the actual utility of the magnetic compass?

16.4 Note that there are two seemingly independent ways of determining the north-south direction at any location on the earth: (1) The direction, at the given location, to the pole star, and (2) the direction, at the given location, of the shortest shadow of a vertical stick, cast as the sun crosses the local celestial meridian.

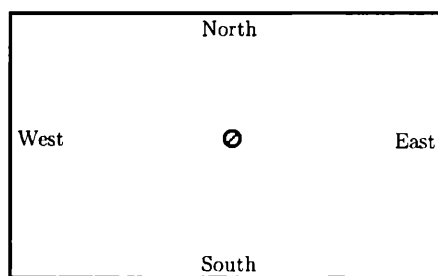
- (a) What significance do you see in the fact that these two entirely different observations give the same direction on the surface of the earth? That is, what

does agreement between these two modes say about axis of rotation regardless of whether one thinks of the earth as rotating relative to the celestial sphere or the entire celestial sphere (including the sun) as rotating around the earth?

- (b) Describe how a *different* arrangement of rotations (one that is *not* actually the case) might have led to disagreement between the two modes mentioned in part (a). (Hint: Consider the possibility of a diurnal motion of the sun different from what is actually observed.)

16.5 Naked eye observations of the sky lead very directly to the concepts of “terrestrial and celestial poles” and “terrestrial and celestial equators.” Describe in your own words how these concepts are formed; in other words, identify the observations that play the key roles and the inferences that are drawn from the observations.

16.6 We are looking vertically down on a sheet of paper with a stick mounted upright at its center (as shown by the small circle). Sketch on the diagram the shadows you would expect to see at your geographic location at the times of day requested below, on the date on which you are doing this problem. In each case choose a reasonable relative length and direction of the shadow, and label it with the letter designating the particular question.



- | | |
|---------------------------|---------------------------|
| (a) Shortly after sunrise | (d) Midafternoon |
| (b) Midmorning | (e) Shortly before sunset |
| (c) Local noon | |

16.7 Draw, for your geographical location, diagrams such as that required in Question 16.6 for the day of the winter solstice, the day of the vernal equinox, the day of the summer solstice, and the day of the autumnal equinox. Be sure to make the four diagrams consistent with respect to both direction and length of shadows at various times of day.

16.8 Suppose that, on a particular day, you make observations of the shadow of a vertical stick from some time in late morning into early afternoon.

- (a) A friend makes exactly similar observations on the same date at a location 2000 miles due east of yours. Do you expect the two records to differ in some substantial way or to be very nearly the same? If you expect them to differ, describe the difference you anticipate. In either case, explain your reasoning.
- (b) Another friend makes similar observations at a point either due south or due north of your location (i.e., on the same meridian), but at the opposite latitude (i.e., on the other side of the equator). How do you expect this person’s observations to differ from yours? Accompany your answer with appropriate sketches of shadows.

16.9 Consider the occurrence of a thin crescent moon:

- (a) How do you account for the dark (unilluminated) portion of the lunar disk?



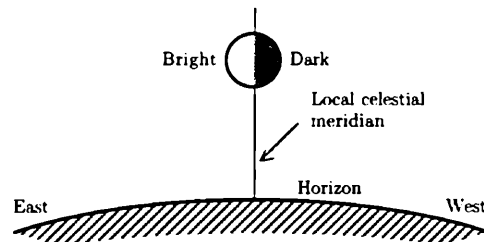
- (b) Can you remember under what circumstances of time of day or night (before or after sunrise, before or after sunset, before or after midnight, before or after noon) you are most likely to notice a thin crescent moon? If you can identify such a time, what relevance does it have to the answer you gave in part (a)?

16.10 (This is an exercise in arithmetical reasoning with the meaning of π and its relation to circular arcs.) The lunar orbit is approximately circular with a radius of 240,000 miles, and the moon shifts eastward through an angle of about 13° in one day.

- (a) Make a sketch of the situation just described, and, explaining each step of your reasoning carefully (do not simply substitute in formulas), calculate the distance the moon travels along its circular arc in one day. [Carry out your calculation in denary (powers of ten) notation.]
- (b) Compare this distance with some distances you know on the surface of the earth and interpret your comparison.

16.11 Suppose you are located in the northern hemisphere, looking south, and you see a moon half-illuminated, as shown, just crossing the local celestial meridian. When we speak of approximate times in the following questions, we refer not to clock hours but to times such as "sunrise," "shortly before or after sunset," "midnight," etc. Be sure to explain your reasoning in answering each question.

- (a) At approximately what time of day or night would you expect to see the configuration shown in the diagram? At approximately what time would this moon have risen? At approximately what time will it be setting?

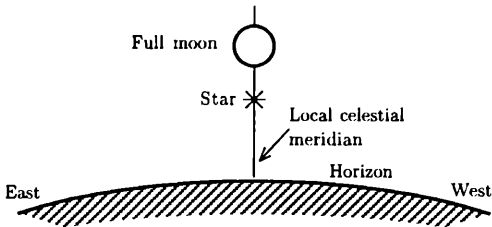


- (b) Suppose you wish to look for the moon approximately 48 hours after the view represented in the diagram. Indicate roughly where in the diagram you would expect to see it, and sketch the approximate shape of the illumination you would expect to see.

16.12 Suppose that on a particular occasion, the full moon and a reference star are seen crossing the local celestial meridian simultaneously, as sketched in the diagram. Explain your reasoning in answering each of the following questions.

- (a) What must be the approximate time of day or night?

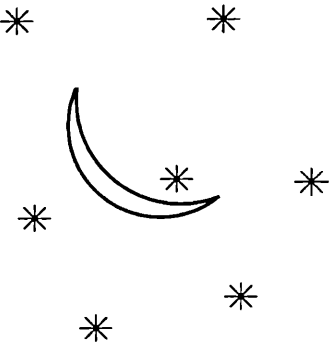
- (b) Sketch what the configuration would look like approximately 3 hours later.
- (c) Sketch the appearance (approximate phase) of the moon and the location of the moon when the same star crosses the local celestial meridian 3 or 4 days later.



16.13 In Samuel Taylor Coleridge’s famous poem “The Rime of the Ancient Mariner,” there occurs the following passage:

“Till clomb above the eastern bar
The hornéd Moon, with one bright star
Within the nether tip.”

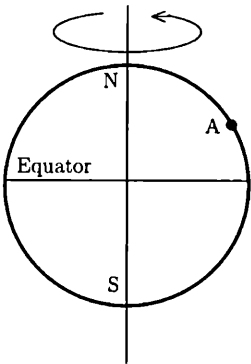
- (a) These lines suggest the accompanying picture of the moon and neighboring stars. Do you see anything wrong with the “star within the nether tip”? Explain your reasoning.
- (b) Is the picture of the moon correct for a northern or for a southern hemisphere viewer? Explain your reasoning.



16.14 What would be your response to a drawing in which a thin crescent moon is shown high in the sky around midnight? Explain your reasoning.

16.15 For simplicity, let us represent the earth as a perfect sphere. The poles and the equator are marked.

- (a) Draw a line that shows the direction of the local zenith point for an observer located at point A, and label this line Z.
- (b) Draw a line representing the horizontal direction for an observer at point A, and label this line H.
- (c) Draw a line that shows the direction along which the observer at A looks to see the North Star, and label this line NS. (Remember that the North Star is extremely far away, a distance many millions of times the radius of the earth.)



- (d) Assuming that the time is that of the vernal equinox and that the sun happens to be located in the plane of the drawing and off to the right, draw a line showing a ray of light from the sun arriving at point A, and label this line S. (Remember that the sun is more than 20,000 earth radii away from the earth.)

16.16 Show in careful detail how you would lead a fellow student to an understanding of the fact that the angular elevation of the North Star above our local horizon is numerically equal to the terrestrial latitude of our point of observation. Be sure to lead the student into defining the term “terrestrial latitude” and carefully drawing and labeling a relevant diagram.

16.17 Suppose you have available a piece of string, a weight, a soda straw, some pins, access to a telephone pole or other form of support, and a protractor. Describe how you might utilize these items to make a rough determination of your present latitude.

16.18 What simple observational evidence can you cite for the fact that the sun is not only farther away from us than the moon but is very *much* farther away?

16.19 Suppose a full moon happens to occur very close to the autumnal equinox. Where along your local horizon would you expect to see this full moon rising? Explain your reasoning with the help of a relevant diagram.

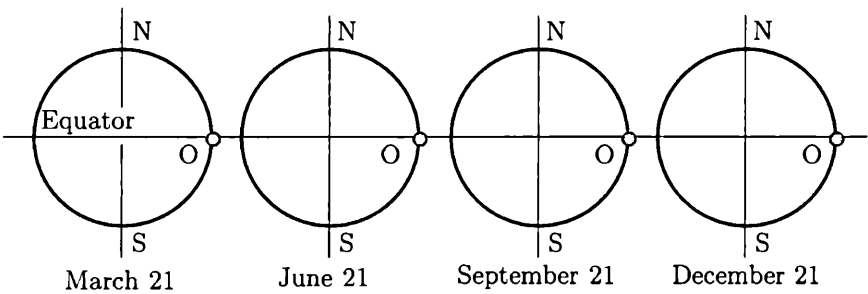
16.20 Let us consider some questions about the location of the sun relative to the local zenith point at noon at various places on the earth. In answering these questions, make use of appropriate diagrams and explain your reasoning.

- (a) Define the term “ecliptic” and prepare to use it when relevant in the following questions.
- (b) Does the sun ever pass through the local zenith point where you happen to live? Is it ever correct to say that “the sun is overhead at noon”? Why or why not?
- (c) At what location on the earth does the sun pass directly overhead at noon at the time of the vernal equinox? Explain your reasoning with the help of a relevant diagram.
- (d) It happens that in San Juan, Puerto Rico, the sun actually does pass through the local zenith point at noon on two occasions: One some days before the summer solstice and the other some days after the summer solstice. Explain this observation with the help of a relevant diagram.
- (e) What is meant by the terms “Tropic of Cancer” and “Tropic of Capricorn,” and what band of behavior of the sun do they bracket on the surface of the earth?
- (f) Suppose you were located close to the north pole. Approximately where would you expect to see the sun at noon on the vernal equinox? Where would you expect to see it at noon a few days later?
- (g) Answer question (f) from the point of view of an observer located simultaneously near the south pole.

16.21 The diagram shows the earth at four different times of the year. An observer O is located at the equator, and the sun is off to the right in the plane of the diagram.

- (a) In each of the four diagrams, draw four rays of light from the sun, incident on the earth, from the point of view of observer O.

- (b) Describe where the sun is located at noon relative to O's zenith point on each of the four dates indicated. That is, is the sun at the zenith? North of the zenith? South of the zenith?

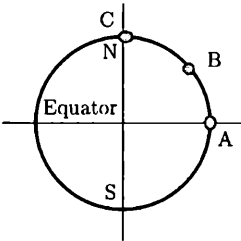


- (c) Describe where the sun rises and sets, from O's point of view, on each of the four dates indicated. That is, does the sun rise north of east? Due east? South of east? etc.
- (d) Describe the relative lengths of daylight and darkness from O's point of view on each of the dates indicated.

Note to the instructor: Question 16.22 can easily be reversed by giving the diagrams and asking for the corresponding dates.

16.22 In the world as we know it, it is an observed fact that the ecliptic and the celestial equator do *not* coincide. For the sake of this question, however, let us imagine a different world in which the ecliptic and celestial equator *do* coincide. In the following questions you are asked to predict how you would expect to see the sun behaving under these circumstances from various points of observation on the earth.

- (a) On the diagram of the earth, show the location of the sun relative to the earth by drawing a set of rays coming to the earth from the sun, which is off to the right. (Remember that the sun is extremely far away relative to the size of the earth.)
- (b) Consider the points of view of observers located at positions A, B, and C: Where would each observer see the sun rising and setting along the local horizon? At what elevation above the horizon would each observer see the sun crossing the local celestial meridian? How would these positions change during the course of the year? Explain your reasoning.



- (c) Describe the relative duration of daylight and darkness for each of the three observers. How would these duration change during the course of the year? Explain your reasoning.

16.23 Roughly what path through the sky (location of points of rising and setting and elevation of point of crossing the local celestial meridian) would you expect the moon to follow at your location if a full moon happened to occur on the winter solstice? If it happened to occur on the summer solstice? Explain your reasoning.

16.24 Suppose the moon happens to be in its third-quarter phase at the time of the vernal equinox. Approximately where along the ecliptic must the moon be located? Explain your reasoning.

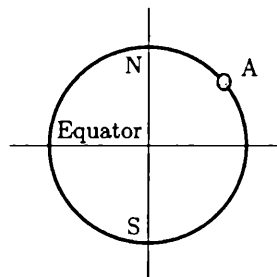
16.25 In this daytime view toward the south at some latitude in the northern hemisphere, the position of the sun is indicated.



At the time being considered, Venus is known to be the “evening star.” Mark on the diagram the position, relative to the sun, at which Venus must be located, and explain your reasoning.

16.26 The very bright star Sirius (sometimes called the Dog Star) follows Orion (the hunter) through the sky. For an observer at a latitude of 48°N , Sirius is observed to cross the local celestial meridian at an elevation of about 22° above the horizon. The diagram shows a cross section of the earth through the local celestial meridian for an observer at point A at latitude 48°N .

- At point A, draw and label the local horizontal and vertical lines.
- Using a protractor, draw and label a light ray arriving from Sirius at point A using the information given above. Mark angles appropriately.
- By adding appropriate lines to the diagram, determine whether Sirius is visible from the north pole. If it is visible, is its highest elevation above the horizon greater or less than 22° ? Explain your reasoning.
- By adding appropriate lines to the diagram, determine whether or not Sirius is visible from a point on the equator. If it is visible, is its highest elevation above the horizon greater or less than 22° ? Explain your reasoning.
- During June, July, and August, Sirius is not visible at all from point A at *night*. How do you interpret this observed fact?

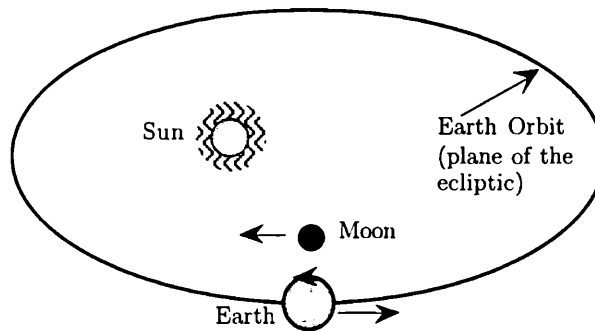


16.27 In our study of elementary terrestrial physics, we gave the name “force” to any interaction that imparts acceleration to a material object, and we gave the name “gravity” to the interaction that accelerates material objects toward the earth.

Describe in your own words how Newton enlarged and extended these concepts in creating a model for the solar system as a whole. Be sure to include the following

ideas as well as others of your own: How do the concepts of “force” and “acceleration” get into the problem at all? What is meant by “extending” the idea of gravity? How does gravity get into the picture at all?

16.28 If we were to look down upon the plane of the solar system from a point above the northern hemisphere of the earth, we would see the earth revolving around the sun, the moon revolving around the earth, and the earth rotating on its axis—all in the counterclockwise direction as illustrated schematically. (The ellipticity of the earth’s orbit and the length scales are greatly exaggerated.)



Given the following definitions: (1) A *solar day* is the time interval for one rotation of the earth relative to the *sun*, i.e., the time interval between the sun’s attaining its highest elevation above the horizon (local noon) at our point of observation. (2) A *sidereal day* is the time interval for one rotation of the earth relative to one of the fixed stars, i.e., the time interval for a star to return to the same position in the sky as it was observed to have on the preceding night. (3) A *synodic lunar month* is the time interval for the moon to return, in its revolution around the earth, to an initial position (or phase) relative to the sun, e.g., the time interval between exact full moons. (4) A *sidereal lunar month* is the time interval for the moon to return, in its revolution around the earth, to the position it initially had relative to some fixed star.

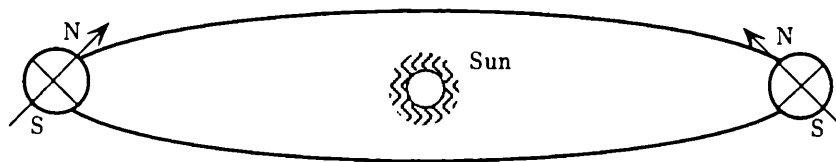
- (a) Note that earth shifts its position in its revolution around the sun over the period of a month as well as over the period of a day. Note also the direction of this shift as well as the direction of the earth’s rotation. In the light of these motions, how would you expect the length of a solar day to compare with the length of a sidereal day? That is, would it be longer, shorter, or of the same length? Explain your reasoning with the help of a diagram.
- (b) It is an observed fact that the earth has a higher velocity in its elliptical orbit around the sun when it is near perigee (closest approach to the sun, occurring in December) than when it is near apogee (greatest distance from the sun, occurring in June). How would you expect the difference between the lengths of solar and sidereal days to change over the course of a year? Explain your reasoning with the help of a diagram.
- (c) With respect to the relation between the synodic and sidereal lunar months, answer the same questions that are asked in parts (a) and (b) with respect to the solar and sidereal days.

16.29 How do you explain the fact that winters are much colder than summers in the northern hemisphere even though the earth is closer to the sun during the winter period than it is during the summer? Use diagrams to assist your explanation.

Note to the student: In the following multiple choice questions, circle the letters designating those statements that are correct. *Any number* of statements may be correct, *not* just one; you must examine each statement on its merits.

16.30 Understanding of a given set of ideas in science is frequently enhanced by carefully thinking through alternative situations, even ones that cannot possibly occur. The following question involves such thinking. We shall deal with the situation sketched in the figure: The earth is visualized as revolving around the sun in such a way that its axis of rotation is always tilted toward the sun. This is physically impossible because maintaining such a condition would require the presence of forces, acting on the earth, that do not exist in the solar system. In the actual situation, the axis of rotation of the earth maintains a very nearly fixed orientation relative to the fixed stars; it does undergo a relatively small motion, called "precession," which is exceedingly slow compared to the annual revolution around the sun, but it plays no significant role in our thinking about the effects that concern us in naked eye astronomy.

Let us contrast the situation illustrated here with what we know to be the case with respect to the revolution of the earth: The tilt of its axis of rotation relative to the plane of the ecliptic, the origin of the seasons, etc.



- (a) Under these circumstances, it would be perpetual daylight at the north pole and perpetual night at the south pole.
- (b) At the latitude at which you are located, days would always be longer than nights throughout the entire year.
- (c) It would always be winter in the southern hemisphere.
- (d) The positions of sunrise and sunset along the horizon (at any given latitude where the sun is observed to rise and set) would not change over the course of the year.
- (e) At any one of the latitudes considered in part (d), the elevation of the sun would always be the same at local noon and would not change as the days go by.
- (f) Instead of there being a fixed polestar, the location of the celestial pole would keep changing during the course of the year.
- (g) The ecliptic (the apparent path of the sun against the field of stars) would not intersect the celestial equator.
- (h) None of the above.

Chapter 17

Learning Objectives

These statements of procedures and learning objectives were given to participants in summer institutes for high school physics teachers at the University of Washington at the start of each subject matter segment of the institute. Most of the participants worked in pairs in an essentially self-paced mode, proceeding through the text and laboratory work of either *PSSC Physics* or *Project Physics* according to their own choice. At the end of a segment, each participant had a conference with a member of the instructional staff, using the statement of learning objectives as a framework.

The conference was originally intended to be an oral examination that determined whether the participant had mastered the material of the unit at a sufficient level to warrant going on to the next subject matter unit. In practice, the conference turned out to be more an opportunity for the participant to synthesize and order his or her own knowledge rather than a test or examination. Some readers of these materials might find that they can be utilized (with suitable modifications and alterations) with students in introductory physics courses at either high school or college level. Some items may be useful as subjects for discussion in collaborative learning groups.

17.1 UNIT 1. KINEMATICS

Participants should start right in on their respective curricula, studying text, working problems, and performing experiments. Working in pairs is encouraged. In general, it is advantageous to have someone to talk with about text and problems and to join forces in the performance of experiments. Learn to weave back and forth judiciously among these various activities according to the intent of the curricula themselves. For example, it is, in many instances, intended that laboratory experiments or observations *precede* reading of the text. Teachers who adhere to the structure and spirit of curricular materials in such respects will be better able to lead students in the manner intended by the progenitors of the curricula.

After you have completed this unit on kinematics, you should be able to do the following things:

- (a) Give clear operational definitions of position s , change of position Δs , instant of time (or clock reading) t , interval of time Δt , average velocity \bar{v} , instantaneous velocity v , average acceleration \bar{a} , and instantaneous acceleration a .
- (b) Translate the verbal description of a rectilinear motion into a graph of position versus clock reading or a graph of velocity versus clock reading (or both).

- (c) Translate given position versus clock reading and velocity versus clock reading graphs into verbal descriptions of the motion represented or into a simulated motion with your own hand or body.

Note: The translations referred to in (b) and (c) can be enhanced and facilitated by use of the sonic range finder coupled to a microcomputer, displaying position, velocity and acceleration graphs of the motion of one's body.

- (d) Work out representative problems at the end of each chapter, explaining your reasoning verbally in each case.
- (e) Describe experiments by means of which you would be able to determine whether observed motions are uniformly accelerated, and be able to determine accelerations experimentally.

The following questions illustrate, but do not limit, some of the questions that might be asked of you in the exit interviews:

- (a) What difficulties do your students have in acquiring an understanding of the concept of "instantaneous velocity," and how would you now go about helping them to overcome these difficulties?
- (b) What is the meaning of the algebraic signs accompanying displacement, velocity, and acceleration in description of rectilinear motion? Explain in your own words how the algebraic signs get into the sequence in the first place. When we adopt a coordinate system and proceed to describe the full history of the rising and falling of a ball thrown vertically upward, does the acceleration of the ball change algebraic sign on the way down relative to the sign on the way up? Why or why not? Explain your reasoning in full detail, indicating how you would lead your own students into understanding what is involved.
- (c) What is the meaning of $v = 0$ at the top of the vertical flight of the ball (or at the extreme end of the swing of a pendulum)? Is the object accelerating when $v = 0$? Is it wise to speak of the object as having "stopped" or "come to rest" at the given instant? Why or why not? What meaning do the terms "stopped" and "come to rest" convey to most people who have not studied physics? How do you propose to handle these matters with your students?
- (d) Sketch a reasonable graph of the position versus clock reading history of several vertical bounces of a ball that has been dropped to the floor from rest at some initial height.
- (e) A car accelerates from rest to a velocity $v = 80$ ft/s in 10 s. During this interval, it has traveled 500 ft. What can you infer about the acceleration in this case? Was the acceleration uniform or nonuniform? Explain your reasoning. Sketch what the v versus t history might have been like. If the distance traveled had been 400 ft instead of 500 ft, what conclusions could you reach about the nature of the acceleration?

17.2 UNIT 2. DYNAMICS

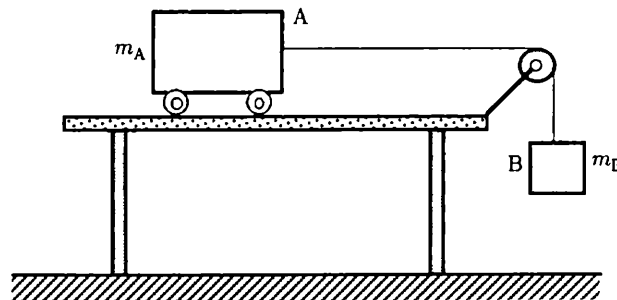
After you have completed the unit on dynamics, you should be able to do the following things:

- (a) Give clear, noncircular, operational definitions of the concepts of “force” and “mass.”
- (b) Define “centripetal acceleration” in circular motion and derive the expression for this acceleration in terms of radius and tangential velocity. (In this connection it also necessary to be able to define “tangential velocity.”)
- (c) Draw correct, well-separated force diagrams of all the objects that interact with each other in a given dynamical situation; describe each force in words (indicating the nature of the force and what object exerts it on what); and identify all third law pairs.
- (d) Describe projectile motion as the superposition of two independent motions, one uniform and the other accelerated. Draw separate force, velocity, and acceleration diagrams for the projectile at any position in its trajectory.
- (e) Work representative end-of-chapter problems in the texts.

The following questions illustrate, but do not limit, some of the questions that might be asked of you in the exit interviews.

- (a) How would you discuss with a student the forces acting on an object resting on a table? (Be able to draw force diagrams for the object, the table, and the earth, indicating the relevant interactions.) What forces are acting on the object? What forces are acting on the table? What happens to these forces as you press down on the object with your hand? What happens to the table when you place the object on it in the first place? (That is, is the table deformed as a result of the force acting on it?) What is meant by the term “passive force” in this context? What happens as you increase the load on the table indefinitely? Is the table deformed or undeformed when you place a sheet of paper on it?
- (b) Suppose a box rests on the floor and you push on it horizontally. Because of the presence of friction, the box does not slide immediately. Describe what happens to the frictional force as you start your horizontal force at zero and increase it until the point of sliding is reached. Draw a force diagram showing all the forces acting on the box and another diagram for the floor, showing the forces the box exerts on the floor. In what ways is this situation similar to that in the preceding question concerning the object placed on a table? In what ways is it different? Can friction be characterized as a “passive force”? Why or why not?
- (c) Draw force diagrams for a cart on an inclined plank when the cart is held at rest and when it is freely accelerating down the plank. In each case also draw a force diagram for the plank. Explain the origin of the word “normal” in connection with the force exerted by the plank on the cart. How would you convince students that this force is indeed normal to the plank and is not directed vertically?

- (d) A ball is placed in a cart or coaster wagon, and the cart is accelerated horizontally in a straight line. Describe carefully what you would see happening to the ball relative to the cart and relative to the ground, as you watched the experiment from the side. What is the point of leading students through such observations? In what sense is it incorrect to say that “the ball is thrown backward”? In what sense might this statement be correct?
- (e) Discuss the floating and sinking of objects in water in terms of forces acting on the chunk of *water* the object will displace when it is immersed and then the forces that must act on the *object* once it has displaced the water. As an outgrowth of this discussion, give a simple nonmathematical justification of the assertion made in Archimedes’s principle, namely, that an object in water is buoyed up by a force equal to the weight of water displaced.
- (f) Suppose that a simple pendulum bob is suspended on a string from the roof of a car and from a rail at the periphery of a merry-go-round. How will the pendulum hang relative to its normal vertical orientation when the car moves at uniform velocity in a straight line? When the car is speeding up? Slowing down? When the car moves around a curve at constant speed? How will the pendulum on the merry-go-round hang when the latter is rotating? Draw separate force diagrams for the bob and the string in each case discussed.
- (g) Suppose you are in an elevator that is in free fall relative to the earth. Describe what you would observe, relative to your frame of reference in the elevator, as you performed various experiments such as (1) letting a ball go from rest, (2) letting the ball go off with some low initial velocity in various directions such as horizontally or up or down, (3) bouncing the ball off the wall, ceiling, or floor, (4) trying to set a pendulum swinging, (5) rotating a bob on a string, and (6) turning over a bucket of water. Make up some other possible experiments of your own.
- (h) Consider the situation shown in the following diagram (you will surely recognize it as a widely used homework problem in the textbooks): The pulley and string have negligible mass; cart A rolls with negligible frictional resistance. The system is released from rest and accelerates as block B falls. We pose the following *qualitative* question: As the system accelerates, how does the force exerted on the cart by the string compare with the weight of block B? Is this force equal to, greater than or less than the weight of the block? Explain your reasoning *qualitatively*, showing how you make use of Newton’s second and third laws to arrive at your conclusion without “solving” the problem analytically through the derivation of algebraic relations.



- (i) Consider a pendulum bob suspended freely from the ceiling or some other support at each of three locations at the surface of the earth: The north pole, the equator, and some intermediate latitude. Let us take the earth to be perfectly spherical (we know this is not actually the case), even though it is rotating. Sketch how the string on which the bob hangs would be oriented relative to a radial line from the center of the earth at each of the three locations if the earth were not rotating and with the earth rotating.
- (j) Suppose we have a bob of mass m on a string and we set it rotating in a circle of radius R that lies in a *vertical* rather than a horizontal plane. Suppose we also manage to keep the tangential velocity v_{tan} constant throughout the circle (it is not important how this is achieved).
 - (1) Consider the bob in two special instantaneous positions: One at the very top of the circle and one at the very bottom. Draw a force diagram for the bob in each case. Define carefully what is meant by the term “centripetal force” and argue that in these two cases, the centripetal force is *not* equal to the force T exerted by the string on the bob. What other force combines with T to make up the centripetal force?
 - (2) Set up algebraic expressions for T at the top and bottom of the circle and interpret them in words, describing what happens to T as v_{tan} is varied from high to low values. How do you interpret the change of sign that T undergoes at the top of the circle in the case of such variation?
- (k) Start with the situation in which a car goes around a curve on an unbanked road (i.e., the road surface lies in a horizontal plane). Draw a force diagram for the car, viewed from the rear, as it is going around the curve, being sure to show the location of the center of the curve in your diagram. What is the origin of the centripetal force that imparts the necessary centripetal acceleration?
 - (1) Now assume that the road is *slightly* banked toward the center of the curve. Draw the force diagram of the car again, and explain why less demand is now placed on the radially directed frictional force on the tires than in the preceding case.
 - (2) Explain what happens to various forces as the angle of banking is increased. How do you identify the angle at which there is no longer any demand placed on the frictional force? What happens if the angle of banking exceeds this value?
 - (3) Suppose we take a fixed angle of banking. Describe what happens on your force diagram as the speed of the car is increased from a very low to a very high value, straddling the optimum speed for which the road is banked.
- (l) Be able to explain and interpret the following experiment that Newton describes in the *Principia*:

I tried the thing in gold, silver, lead glass, sand, common salt, wood, water, and wheat. I provided two equal wooden boxes. I filled one with wood, and I suspended an equal weight of gold (exactly as I could) in the center of oscillation of the other. The boxes, hung by equal threads of 11 feet, made a couple of pendulums perfectly equal

in weight and figure, and equally exposed to the resistance of the air. Placing the one by the other, I observed them to [swing] together forwards and backwards for a long while with equal vibrations. And therefore the quantity of matter [inertial mass] in the gold was to the quantity of matter in the wood as the action of the motive force [gravitational mass] upon all the gold to the action of the same upon all the wood; that is the weight of one to the weight of the other.

What was the point of this experiment? What did Newton observe? What inference is drawn from the results?

- (m) Explain the basis for our belief that in the solar system, the earth and planets revolve around the sun rather than all the other members revolving around the earth.

17.3 UNIT 3. MOMENTUM AND ENERGY

After completion of this unit, you should be able to do the following things:

- (a) Give clear and detailed operational definitions of all the new terms that have been introduced such as “impulse,” “momentum,” “work,” “kinetic energy,” and “potential energy.”
- (b) Describe what happens in various commonly observed phenomena in terms of these new concepts. For example: describe, in terms of impulse delivered, momentum changes, work done, and various energy changes, what happens when a ball is thrown vertically upward and eventually ends up on the ground. Describe what happens when carts with spring bumpers collide and bounce apart. Describe what happens when a box is pushed along the floor so that it accelerates in the presence of friction, etc.
- (c) Define the concept of “center of mass” and describe its role in phenomena involving momentum and energy transformations.
- (d) Be able to solve both qualitative and quantitative end-of-chapter problems.

The following questions illustrate, but do not limit, some of the questions that might be asked of you in the exit interviews. (Be forewarned that these are sophisticated questions that are poorly handled in virtually all existing introductory texts.)

- (a) Describe how you would lead your students, through appeal to common and easily observable phenomena, to distinguish between “temperature” and “heat,” realizing, to begin with, that many students use these terms synonymously.
- (b) Suppose you push a box along the floor at uniform velocity with a horizontal force P in the presence of friction. The frictional force f must be equal in magnitude and opposite in direction to P . In a displacement Δx you have done an amount of work $P\Delta x$. What has happened to this amount of work? Where, that is, has the energy gone? Is it correct to say that a negative amount of work ($-f\Delta x$) has been done by the frictional force and therefore zero net work has been done on the block? Why or why not? (Remember that in physics we wish to define “work” as a form of energy put into, or taken out of, a system that has been specified—not just any occurring product of force and displacement.)

- (c) In deriving the work-energy theorem, starting from $F_{\text{net}} = ma$, the displacement Δx that appears in the derivation must be that of the center of mass of the particle being accelerated. Thus the quantity of work done under these circumstances is the net force integrated over the displacement of the center of mass of the particle. Suppose we now proceed to compress a spring. Is the work done in compressing the spring the force exerted on the spring integrated over the displacement of the center of mass of the spring? What *is* the correct calculation? If we compress a gas in a cylinder by displacing a piston, is the work done on the gas equal to the force exerted by the piston integrated over the displacement of the center of mass of the gas?
- (d) Suppose you stand on roller skates and push yourself away from a wall, leaving contact with the wall with kinetic energy of your entire body. Where has this energy come from? Did it come from work done by the wall on your body? Why or why not? Answer the same questions with respect to the situation in which you jump vertically upward. What is the source of the discrepancy we have run into in questions (c) and (d)? What are some basic limitations we must place on applicability of the work-energy theorem derived from $F_{\text{net}} = ma$?

17.4 UNIT 4. ELECTRICITY AND MAGNETISM

After you have studied the text material on electricity and magnetism and have done representative homework problems, you should be able to give clear and specific answers to the following questions in exit interviews.

- (a) In operational terms, under what circumstances do we describe an object as being “electrically charged”? What is meant by the term “electrical charge”? What can one say about what electrical charge “is”? What observations lead us to visualize electrical charge as *transferable* from one object to another?
- (b) In electrostatic situations, what observations lead us to characterize some substances as “conductors” of electric charge and other substances as “nonconductors”?
- (c) What observations lead to characterizing two objects as carrying “like” charges and to the assertion that “like charges repel”? That is, does introduction of the term “like” stem from arbitrary definition or from experimental observations?
- (d) What is meant by the term “unlike charges”? Use of only two terms (like and unlike, positive and negative) implies that there are no more than two varieties of electrical charge. What leads us to accept this idea? (What experimental observation would force us to recognize a third variety if it should turn up?)
- (e) Note that electrical charge is *not* a material substance and cannot be measured by its mass or any other such “tangible” property. Through what sort of experimental observations is it possible to assign numerical values to “quantity of electrical charge”? What observable effects, that is, can be used to help quantify electrical charge?
- (f) State Coulomb’s law clearly and precisely, defining all terms occurring in the algebraic statement, and indicating to what situations this law is directly applicable as well as at least one or two situations in which it would *not* be directly

applicable. (Is the law directly applicable, for example, to the case of two charged rods attracting or repelling each other?)

- (g) It is an observed fact that a charged object carrying either variety of electrical charge attracts any uncharged object. What explanation is invented for this occurrence? What prior observations and experiences motivate the construction of this model (the model to which we give the name “polarization”)?
- (h) Consider the following situations in which there is a clear interaction between an electrically charged object and an uncharged object: (1) A charged rod attracts an uncharged pith ball or a bit of paper; (2) a charged rod charges, on contact, a pith ball covered with a conducting coating; (3) the leaves of an electroscope separate when a charged rod is brought near the electroscope without actually making contact; and (4) an electroscope is charged by induction.
 - (1) In each of these preceding cases be sure that you are able to describe, in terms of any one of the three following models and with the aid of clear diagrams, the interactions that take place: (i) Negative charge is taken to be stationary while positive charge is free to move; (ii) positive charge is taken to be stationary while negative charge is free to move; and (iii) both varieties of charge are free to move. Argue, in your own words, that at the level of macroscopic observations, it is impossible to confirm or to rule out any of these three models.
- (i) Consider the following sets of interactions among various objects—interactions to which we have already given different names: (1) Electrostatic interactions of charged bodies with each other and with uncharged bodies, (2) magnetic interactions of permanent magnets with each other and with unmagnetized pieces of iron such as nails or paper clips, and (3) gravitational interactions among all material bodies in the universe.
 - (1) Describe as many observations as you can that show that, despite many similarities, there are many profound differences among the three types of interaction, and argue that these differences provide the justification for recognizing three different phenomena and therefore for creating three different names.
- (j) If, in your studies, you introduced the term “electricity” initially in connection with what we call “electrostatic interactions,” describe experiments and observations that justify the conclusion that the phenomena associated with what we call “electrical batteries” and household “electrical outlets” are related to electrostatic interactions in such a way that it is indeed legitimate to invoke the term “electrical” for all these seemingly very different phenomena. If, on the other hand, you started with what we call “current electricity,” describe observations and experiments that reverse the preceding line of argument and justify using the word “electrical” in connection with interactions among objects that have been rubbed with other materials.
- (k) Be able to sketch electrostatic force patterns (i.e., “electrical lines of force”) in the neighborhood of various simple configurations such as: A point charge, a line charge, relatively close to a charged plane, a uniformly charged sphere,

charged capacitor plates, two like point charges, and two unlike point charges. (Be sure that you know the convention for establishing the *direction* of the field at any point, not just the overall pattern.)

- (l) Be able to account for the fact that except for small fringe effects near the edges, the electrical field between capacitor plates is confined to the region between the plates with zero field outside, while the electrical field of a single charged plate extends to great distances (relative to the size of the plate) on either side.
- (m) Be able to sketch the magnetic force patterns (i.e., “magnetic lines of force”) around such simple magnetic and electromagnetic configurations as a single magnet, between like and unlike magnetic poles, around a current-carrying wire, around a coil of wire, and around a solenoid. (Be sure that you know the convention for establishing the *direction* of the field at any point, not just the overall pattern.)
- (n) Be able to predict the direction of force on an element of current-carrying wire at any location in a magnetic field of given direction.
- (o) Be able to indicate the direction of the force (i.e., the “Lorentz force”) that would act on either a positive or negative electrically charged particle moving in any given direction at any specified location in a magnetic field.
- (p) Two parallel wires carrying electric current in the same direction are observed to attract each other. Following Ampere, we call this interaction “electromagnetic.” Since the wires are connected to a battery or to some other electrical source, the interaction might have been electrostatic. How do we know that it is *not* electrostatic? That is, what observations and experiments compellingly indicate that the interaction differs completely from electrostatic interactions?

17.5 UNIT 5. WAVE PHENOMENA

I. After you have studied the behavior of waves in one dimension (e.g., waves on a stretched string and on a spring such as a “slinky”) you should be able to do the following:

- (a) Describe, in your own words, without use of any new technical terms, what is meant by the terms “wave pulse” and “wave train.” In developing your description, make clear what behaviors distinguish wave motion from the ordinary particle motion studied earlier. Is particle motion still present in the wave phenomena you are now describing? If so, in what way? (Be sure to sketch relevant pictures in the course of your description.)
- (b) Describe clearly and precisely in ordinary language (i.e., with a minimum of technical terms) and with the help of relevant diagrams the differences between transverse waves on the one hand and longitudinal waves on the other.
- (c) Define the term “propagation velocity V ” of a wave pulse and describe what factor or factors determine or affect this velocity in transverse and longitudinal pulses in various media.

- (d) Define the term “particle velocity” in a wave pulse and point out the relevant motions in actual transverse and longitudinal pulses that you initiate on strings and slinkies. How is the particle velocity directed (relative to the direction of wave propagation) in the case of a longitudinal compression pulse? In a longitudinal rarefaction pulse?
- (e) Describe, sketch, and interpret graphs that might be used to represent propagating *transverse* wave pulses. (Note that there are two possible variables for the abscissa: Clock reading t and position x along the medium. There are two possible ordinates: Transverse particle displacement y from equilibrium position and transverse particle velocity v_y in the y direction. You should be able to use and interpret any of the possible representations.) Interpret the meaning of the algebraic signs that accompany the variable y (i.e., how do these signs originate and how are they to be interpreted physically?). Define the term “amplitude” in each representation.
- (f) Given a stretched string, describe how you would generate (1) a transverse pulse with only positive deflection, i.e., only a positive phase; (2) a transverse pulse with only negative deflection, i.e., only a negative phase; and (3) a transverse pulse with both positive and negative deflections, i.e., with both positive and negative phases. (In your description, indicate what motion you would actually execute with your hand at one end of the string.)
- (g) Describe and interpret graphs that might be used to represent propagating *longitudinal* wave pulses. What possibilities are there for the ordinate? (Note that in addition to particle displacements and velocities, there is the possibility of defining a “density” of the medium that could serve as a useful variable.) Interpret the meaning of the algebraic signs that accompany the variables you have chosen as ordinates for your graphs. How do these signs originate and how are they to be interpreted physically? What is meant by the terms “compression” and “rarefaction”? Define the term “amplitude” in each representation.
- (h) Given a stretched coil spring (say, lying on the table), describe how you might generate (1) a longitudinal pulse with only a positive phase, (2) a longitudinal pulse with only a negative phase, and (3) a longitudinal pulse with both positive and negative phases. (In your description, indicate what motion you would actually execute with your hand at one end of the spring.) Now return to the description you gave in item (f) above, and identify the very significant differences between what happens in the case of transverse waves on the one hand and longitudinal waves on the other when it comes to generating pulses having only one phase, positive or negative.
- (i) Describe what happens when transverse and longitudinal wave pulses traveling in opposite directions “collide.” How does the behavior differ from what happens when particles collide?
- (j) Define what is meant by “superposition” of wave shapes, and describe what happens when superposition occurs with either transverse or longitudinal pulses. (You should be able to describe superposition either in terms of particle displacements or particle velocities.)
- (k) Define, using words and diagrams, “constructive interference” and “destructive interference” in terms of superposition of wave pulses.

- (l) Given the shape of a transverse or longitudinal pulse incident at a fixed or at a free boundary, be able to predict the shape of the reflected pulse. (You should be able to deal with pulses of asymmetric shape, either positive or negative, either compressions or rarefactions. You should be able to make such predictions through visualizing the reflected pulse as emerging from the fictitious region beyond the boundary and having a shape such that, when superposed on the incident pulse, the layer of medium at the boundary maintains its fixed or free character. Maintaining the character of the boundary is called “satisfying the boundary conditions.”)
- (m) Describe what is meant by a “continuous wave train” in contrast to a single pulse. Describe what is meant by a “periodic” wave train. Describe what is meant by a “sinusoidal” wave train. Describe how you would generate transverse and longitudinal periodic wave trains in the media with which you have worked. Define and illustrate with appropriate diagrams what is meant by “constructive” and “destructive” interference of periodic wave trains.
- (n) With the help of appropriate diagrams, define the terms “frequency, f ” and “wavelength, λ ” of a periodic wave train. Then, reasoning *arithmetically* from the definitions, establish the relationship among the three quantities V (velocity of propagation), f , and λ . (In other words, you should be able to reason out the relationship whenever you need it rather than memorizing it as a rigid formula.)

II. These learning objectives concern the behavior of straight and circular wave pulses (not wave trains) generated in a ripple tank. After studying the behavior of such pulses, you should be able to do the following:

- (a) Define the concepts “ray” and “wave front,” using diagrams as well as words, and illustrating the concepts in the cases of both straight and circular pulses.
- (b) Using appropriate diagrams, define the concepts “angle of incidence” and “angle of reflection” for both straight and circular pulses incident at a straight rigid barrier. Define “normal” and “glancing” incidence. Be able to sketch these angles in wave front as well as ray representations. Be able to sketch what happens as the angle of incidence is increased from normal to glancing.
- (c) Suppose a straight wave pulse propagating in a region of deeper water (higher propagation velocity) is incident at a straight interface with a region of shallower water (lower propagation velocity). Sketch separate ray and wave front diagrams showing the incident, transmitted and reflected pulses. Sketch such diagrams for the case in which the situation is reversed and the incident wave pulse arrives in the shallower region. In connection with your diagrams, define the concept “angle of refraction” (or “angle of transmission”).
- (d) In each of the instances and diagrams arising in item (c), be able to sketch how the angle of refraction changes as the angle of incidence is varied from normal to glancing.
- (e) On the basis of your observations with circular wave pulses, be able to sketch what happens in reflections from a straight rigid barrier; i.e., show how the reflected wave changes as you move the center of the incident wave closer to, or farther from, the barrier.

- (f) Sketch what happens when circular pulses are incident at a straight refracting interface, making diagrams that show different distances of the center of the circular pulses from the interface.
- (g) In moving an object, say a pencil, through a ripple tank, describe the circumstances in which a bow wave is *not* formed by the moving pencil and the circumstances in which a bow wave *is* formed. Explain how the two situations differ.

III. These learning objectives concern the behavior of periodic wave trains in a ripple tank:

- (a) Sketch the wave front patterns observed in the ripple tank when straight wave trains are incident at a straight rigid barrier at different angles of incidence.
- (b) Sketch the wave front patterns observed when straight wave trains are incident at a refracting interface at different angles of incidence. (Be able to do this for incidence in both the deeper and the shallower water regions.)
- (c) A straight wave train arrives at normal incidence to a straight barrier that is shorter than the length of the wave fronts (i.e., the unimpeded portion of the wave front can propagate past the barrier while part of the wave front is blocked). Sketch the pattern to be observed in the region beyond the barrier for the cases in which the wavelength λ is very short relative to the length of the barrier, very long relative to the length of the barrier, and of intermediate length.
- (d) A straight wave train arrives at normal incidence to a straight barrier that contains an opening of width D , and the waves are blocked except for passage through the opening. Sketch the pattern of wave fronts transmitted through the opening for different ranges of the ratio of wavelength to opening width λ/D , i.e., for small, large, and intermediate values of λ/D .
- (e) Explain why the wavelength of a wave train changes when the train is transmitted through a refracting boundary while the frequency of the train remains unchanged.
- (f) Circular wave trains of wavelength λ , from two point sources running in synchronism a distance d apart, form an interference pattern. Sketch the pattern for different values of d at a fixed value of λ and for different values of λ at a fixed value of d . Identify the regions of constructive and destructive interference in each pattern. Selecting any arbitrary location in any one of the patterns, be able to say how many wavelengths the wave front from one of the sources leads or lags the corresponding (simultaneously emitted) wave front from the other source.
- (g) In the light of the ideas that have been developed, define the concepts "refraction," "diffraction," and "interference" and distinguish among them.

IV. The following learning objectives involve the experience of transferring, in the abstract, what has been learned about waves on coil springs and in ripple tanks to the phenomenon of sound waves:

- (a) In the light of what you have seen of the generation and behavior of compression and rarefaction pulses on a coil spring, sketch what you visualize might be happening to the air in a tube when a piston or diaphragm moves rapidly back and forth at one end.
- (b) What variables (ordinate and abscissa) might you use to make a graph of a sound pulse?
- (c) Is it possible to make a sound pulse having only a compression and no rarefaction phase by moving the piston inward and returning it to its initial position? Why or why not? How does this situation compare with the generation of pulses on the coil spring?
- (d) How would you imagine the interference of sound waves to take place? Suppose you had two audible point sources of sound (analogous to the situation with two point sources in the ripple tank), e.g., two tuning forks sounding in unison. How would you go about finding regions of constructive and destructive interference?
- (e) Use what you have said in items (d) and (l) in Section I above to predict how the compression and rarefaction phases of sound pulses would be reflected from a rigid wall.

17.6 UNIT 6. THE NATURE OF LIGHT

After studying the material concerning the model for the behavior of light that is accepted in classical physics, you should be able to do the following:

- (a) Describe experiments in which the behavior of light can be consistently accounted for on the basis of a corpuscular (i.e., particle) model.
- (b) Describe experiments in which a corpuscular model fails to provide a consistent model for the behavior of light and explain in what way the corpuscular model fails.
- (c) Describe in what way the wave model is successful where the corpuscular model fails and therefore why the wave model is ultimately accepted.
- (d) Describe how the wave model accounts for the observed fact that the center of the Newton's rings experiment (where the thickness of the air film between the two pieces of glass is very much smaller than the wavelength of light) is dark rather than bright.
- (e) Describe experiments that indicate light waves to be transverse rather than longitudinal, even when one still does not know the wave to be electromagnetic in nature.
- (f) Describe commonly observed phenomena that hint (not necessarily prove) that light might be intimately connected with electric and magnetic effects on the microscopic level of the structure of matter.

Chapter 18

Term Paper Assignments

Term paper assignments can have a variety of desirable goals, and each teacher will emphasize different priorities. The following sample assignments have the principal goal of cultivating aspects of “scientific literacy” such as those defined in Chapter 12 of Part I.

Vaguely stated term paper assignments are not, in general, very effective for students in introductory courses. Without at least some guidance toward ways of focussing discussion and development, only a very few students are likely to write papers that penetrate as deeply as many are fully capable of doing. The problem for the instructor is to provide enough guidance to help focus attention on significant issues without imposing excessive constraint or specifying desired conclusions. Experience with such assignments indicates that guidance is most fruitful when it marks out degrees of specificity in the discussion being called for but allows substantial range of choice as to examples to be used and ideas to be analyzed.

It is also fruitful to lead students into describing and assessing their personal learning experiences in the given context. At this stage very few students, if any, have had the opportunity to assess what it means to have learned and understood a significant range of scientific ideas, and even fewer have had the opportunity to describe a learning experience of their own in their own words. Given such opportunity for reflection and assessment, many students articulate penetrating insights that do not emerge in other circumstances.

Following are a few examples of term paper assignments the author has used, with varying degrees of success and with various student populations. Most have to do with aspects of the Newtonian synthesis, but one has to do with the concept of the “electron.” The intent here is not to promulgate specific assignments but rather to provide at least one model for the construction of term paper assignments that enhance student response and performance. A teacher should, in the final analysis, generate his or her own assignments geared to the level of readiness and the vocabulary of the students.

18.1 EXAMPLE 1

Write a coherent essay, approximately 1000 words in length, describing the evolution of the law of universal gravitation as we studied it in its historical context and discussing the alteration it represented in contemporary points of view toward terrestrial and celestial phenomena. Include examples of calculations that you yourself

can make as a result of what you have been learning. You might, if you wish, include a discussion of the contribution made by Cavendish (the experiment he described as “weighing the earth” and which we interpreted as measuring the universal constant G). If you elect to do so, be sure to indicate when, in the historical sequence, this work was done.

In the course of your discussion, show that you understand the meaning of the term “empirical information” in connection with Kepler’s laws and the relevance of this idea to the algebraic development of the gravitation law. Also show awareness of the distinction between inductive and deductive reasoning and point out specifically, as you go along, what kind of reasoning is being described at any given point.

In illustrating calculations you yourself can make, you are free to choose problems from the textbook or to make up a problem or problems of your own (special credit will be given for the latter mode). You might calculate the tangential velocity of a satellite in near-earth orbit and compare it with values that have been quoted in newspapers or magazines. In any case, in working out any problem, explain all steps as an author would in a textbook, and interpret your results. Worked-out problems should, however, be incorporated smoothly in the story line of your paper, not presented as abrupt, isolated entities.

18.2 EXAMPLE 2

The following quotation is from *Patterns of Discovery* by N. R. Hanson (Cambridge University Press, Cambridge, 1958).

[The formula] $F = Gm_1m_2/R^2$ did remarkable work in the ‘Principia’. . . for the law unified the laws of Kepler and Galileo into a powerful pattern of explanation—one of the most powerful in the history of physics. For Newton the law . . . did not simply ‘cap’ a cluster of prior observations: it did not summarize them. Rather it was discovered as that from which the observations would become explicable as a matter of course. Newton was not an actuary who could squeeze a functional relationship out of a column of data; he was an inspired detective who, from a set of apparently disconnected events (a bark, a foot print, a ‘faux pas,’ a stain) concludes, ‘The game keeper did it.’ No one less than a Newton, given the laws of Galileo and Kepler, observations of the lunar motion, the tides, and the behavior of falling bodies, could infer that $F = Gm_1m_2/R^2$. This law organizes and patterns all those things, and others as well, but nothing incompatible with any of them.

The conceptual situation is not unlike this: Novel mathematical theorems are encountered which, besides being individually surprising, do not seem to fit together as a system. . . . Philosophers sometimes regard physics as a kind of mathematical photography and its laws as formal pictures of regularities. But the physicist often seeks not a general description of what [he or she] observes, but a general pattern of phenomena within which what [is observed] will appear intelligible. It is thus that observations come to cohere systematically. . . .

Write a paper of approximately 1500 words using the quotation above as a point of departure for discussion of what you learned in studying the Newtonian synthesis.

(Note that the Newtonian synthesis includes the laws of dynamics as well as the law of gravitation.) You might want to consider and deal with at least some of the following questions. It is up to you to decide what to write; it is not possible to cover all the questions suggested.

- (a) Are you surprised by any of Hanson's remarks? Do you agree or disagree with him? (There is ample room for disagreement. You are not obligated to accept everything he says.) Why do you think Hanson says that the law of gravitation did "not simply 'cap' a cluster of observations"? In what sense is it reasonable to say that Newton was an "inspired detective" rather than an "actuary"? (If need be, look up the meaning of the latter word.)
- (b) In what sense is the word "explanation" used in the context under consideration? What do you see to have been "explained" by the Newtonian synthesis? What do you see as *not* explained? Has $F = ma$ been "explained"? What relevance do you see in Newton's famous remark: "But hitherto I have not been able to discover the cause of those properties of gravity from phenomena [i.e., observation and experimentation], and I frame no hypothesis. . . . To us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies and of our sea." (It should be noted that to Newton and his contemporaries, the term "hypothesis" referred to mystical or occult ideas that were to be avoided in natural philosophy. Thus it carried a pejorative connotation—not a favorable and respectable one as it does at the present time.)
- (c) How do *both* m_1 and m_2 get into the gravitation formula? What is the connection between "mass" in this context and inertial mass m in $F = ma$?
- (d) Look up the meaning of the words "inductive" and "deductive" as used in connection with scientific reasoning and list at least three or four specific instances of each type of reasoning in the sequence you studied. What do these instances have to do with Newton's "detective work"?
- (e) Is a philosopher wrong in regarding physics as a "a kind of mathematical photography"? What do you think Hanson means in the sentence that contains this metaphor? What is meant by "observations come to cohere systematically"? Do you think that the basic laws or formulas "come to cohere systematically" in the same sense as do the observations to which the formulas apply? Why or why not?
- (f) Has your study of this episode in the history of ideas altered or expanded any of your notions of the nature of science? If so, in what way? When did the alteration occur? What have you learned aside from "physical facts"? What is your present view of what it was that Newton learned?

18.3 EXAMPLE 3

[The Austrian physicist Ernst Mach, who was a professor of physics at the University of Prague, was one of the great physicist-philosophers of the 19th century. In 1883, in his book titled *The Science of Mechanics*, he subjected the Newtonian theory to

a most sophisticated and searching criticism. With judgments, insights, and perspectives inherited from the two hundred intervening years of scientific thought, he devastatingly analyzed Newton's sometimes circular definitions, fallacious justifications of the concepts of absolute space and time, and repeated appeal to "unscientific hypotheses" (which Newton professed to avoid but slipped into unwittingly). Mach provided an important part of the criticism and analysis that ultimately led to Einstein's reexamination of the foundations of the entire theoretical structure and his formulation of the theory of relativity. We shall not concern ourselves with Mach's critical analysis but shall use as a keynote some of his remarks about the Newtonian synthesis in general. The following passage is taken from the English translation of the second German edition of *The Science of Mechanics*. The Open Court Publishing Company, Chicago, 1893.]

But in addition to the *intellectual* performance [in the *Principia*], the way to which was fully prepared by Kepler, Galileo, and Huygens, still another achievement of Newton remains to be estimated, which in no respect should be underrated. This is an achievement of imagination. . . . Of what nature is the acceleration that conditions the curvilinear motion of the planets . . . ? Newton perceived, with great audacity of thought, and first [according to his own reminiscences] in the instance of the moon, that this acceleration differed in no substantial respect from the acceleration of gravity so familiar to us. It was probably the principle of continuity, which accomplished so much in Galileo's case, that led him to his discovery. He was wont . . . to adhere as closely as possible, even in cases presenting altered conditions, to a conception once formed, to preserve the same uniformity in his conceptions that nature teaches us to see in her processes. . . . The motion of the moon thus suddenly appeared to him in an entirely new light, but withal under quite familiar points of view. . . . The new conception was attractive in that it embraced objects that previously were very remote, and it was convincing in that it involved the most familiar elements. This explains its prompt application in other fields and the sweeping character of its results. . . . Thus an amplitude and freedom of physical view were reached of which men had no conception previously to Newton's time.

Write a paper of approximately 1500 words using the quotation above as a point of departure for discussion of what you learned in studying the Newtonian synthesis. (Note that the Newtonian synthesis includes the laws of dynamics as well as the law of gravitation.) You might wish to amplify some of Mach's comments with specific illustrations. You might want to consider and deal with at least some of the following questions. It is up to you to decide what to write; it is not possible to cover all the questions suggested.

- (a) In what sense is the term "synthesis" used in this context? What was it that Newton "synthesized"? What aspects did the synthesis not include? Be sure to give specific examples.
- (b) In what sense did the motion of the moon appear "in an entirely new light"? Did you have a similar personal experience with aspects of the ideas being studied? If you did, try to identify and describe the nature of this experience.

- (c) What was the “principle of continuity that accomplished so much in Galileo’s case”? (Galileo, for example, argued that the restrained behavior of the ball rolling down the inclined trough in slower, more readily measurable motion, should be related to its behavior in free fall.) What arguments based on such uniformity in nature have you encountered up to this point? Give several specific illustrations of Newton’s use of arguments based on continuity and uniformity among apparently disparate natural phenomena.
- (d) If you have an interest in intellectual history, illustrate and expand on the consequences alluded to in the last sentence of the quotation.

A note of information: Newton devoted considerable time and effort to accurate measurements of the period of a pendulum consisting of a hollow housing into which he placed various materials differing in chemical nature and composition. From these measurements he calculated the value of acceleration due to gravity, g , showing it to be independent of the nature of the materials. Why do you think he did these experiments? You might find this episode relevant to dealing with some of the questions posed above.

In writing this paper, you are not being asked to become a historian of science and discover lines of reasoning by which Newton himself arrived at various insights. Such specific details are hidden for the most part even from professional historians. Newton published in what was to him the final, polished, convincing form and allowed very little of his “private science” to become known to others. Your task is simply to *interpret* and *illustrate*, thereby deepening your own understanding of this dramatic incident in our cultural heritage.

18.4 EXAMPLE 4

The following quotation is taken from *The Mechanization of the World Picture* by the Dutch historian of science E. J. Dijksterhuis (Clarendon Press, Oxford, 1961).

. . . [P]eople had always speculated about the movements performed by material bodies under the influence of internal or external causes; for this purpose they had used various terms, which, because they also occurred in everyday speech, appeared quite clear, but which in reality were not defined sharply enough to be simply or safely employed in scientific discussions. Scientists had spoken of gravity, levity, force, power, velocity, resistance, tendencies, sympathy, antipathy, impetus, quantity of motion, mass, centrifugal force, . . . and the force of an impact, without ever having adequately defined any of these concepts. Without formulating them explicitly, scholars had started from certain general notions, which had been borrowed from day-to-day experience and therefore appeared to be evident, but afterwards all these notions were found too inadequate for an exact treatment of the subject to be based on them. Gradually, indeed, doubt had arisen as to the correctness of the . . . dynamics founded on these notions; in particular a new notion of inertia had replaced the old one. But while this was going on, other conceptions from ancient dynamics, specifically the proportionality of force and velocity, had been preserved in full, and scientists had always omitted to define accurately the numerous terms they used. . . .

It was Newton's task to create order in this chaos of terms and notions. The best method would have been that of Hercules cleansing the Augean stables, i.e., radical rejection of the old and subsequent reconstruction from the bottom. In this case it would have meant placing mechanics on a new foundation with the aid of sharply defined terms, preferably not taken from everyday speech, so that they were not yet charged with misleading associations. But science in its actual development is not accustomed to heed such semantic ideals: any man who tries to reorganize it has himself grown up in the world of thought he wishes to reform, thinks in its concepts, and speaks in its terms. . . .

Use this quotation as a point of departure for discussion of your own learning experience in studying Newtonian mechanics. Fulsome adulation of Newton and uncritical acceptance of every remark of a commentator are unnecessary. There is ample room for agreement and disagreement, for modifications of point of view, and for acknowledgment of ignorance. Think through your ideas and your experience carefully, and, drawing on your present level of knowledge, present them with some courage of conviction, using specific illustrations to support your statements.

You might wish to consider and enlarge on some of the following questions.

- (a) Has your own view of the meaning of terms such as "velocity," "acceleration," "force," "mass," "inertia," and "gravity" changed in the course of this study? If so, at what point and in what ways?
- (b) What do you see to be the role of language and the influence of everyday speech in clouding or illuminating formation of scientific insights? Do you think it might have been possible for Newton to have "cleansed the Augean stables"? If so, how might he have done so? (Be sure to identify the origin, and explain the meaning, of this metaphor.)
- (c) In the light of your experience so far, what problems can you begin to anticipate in connection with introduction of terms such as "energy," "work," "heat," and "power" in forthcoming discussions?
- (d) In the light of your own study of this episode in the history of science and ideas, do you agree that Newton brought "order out of chaos"? Why or why not?
- (e) If you have taken courses in English literature, have you encountered any analogous problems in connection with interpretation of a literary work?
- (f) Have you encountered references to Newton or the Newtonian synthesis in any of your courses in literature or history? If so, you might wish to connect some of these references with what you have to say in this paper. Such connections deserve a bonus in grading.

18.5 EXAMPLE 5

Note to the instructor: The first part of this assignment should be made prior to study of the Thomson experiment on charge-to-mass ratio of entities in the cathode beam as outlined in Chapter 10 of Part I.

Write a brief (one-page) paper addressed to the following questions: What does the term “electron” mean to you at the present time? How do we come to know about such an entity? What evidence is there for its existence? What are some of its properties? (Submit one copy of the paper to your instructor and keep one copy for your own use in connection with the second part of this assignment.)

Note to the instructor: The second part of the assignment should be made after study of Thomson’s experiments on the cathode beam as outlined in Chapter 10 of Part I.

In the light of your study and interpretation of the Thomson experiment, address the questions about the concept of “electron” that were raised in the first part of this assignment. Compare your present view of the concept with that which you held prior to this study. In what ways, if any, have your insight and understanding changed? Describe in detail, using specific examples. What do you believe it means to “understand” such a scientific concept?

PART III

Introduction to the Classical Conservation Laws

Preface to Part III

With only a few exceptions, presentations of the energy concepts in introductory physics texts suffer from certain “infelicities” or are somewhat misleading. The origin of the difficulty resides principally in how the Work-Kinetic Energy Theorem (W-KET) is derived and generalized. Although this theorem, derived from Newton’s Second Law, applies only to displacement of a point particle or the displacement of the center of mass of a system of particles, this profound restriction is not adequately emphasized, and applications to extended, deformable bodies are either made arbitrarily without justification or are made incorrectly.

In calculating the work done in compressing a spring, for example, it is asserted that one must integrate over the displacement of the end of the spring rather than the displacement of the center of mass, even though work, according to the W-KET, is to be calculated over displacement of the latter. Justification for this sudden change is rarely, if ever, given explicitly.

In dealing with jumping vertically upward or pushing oneself away from a wall while standing on roller skates, it is frequently stated or implied that the normal force exerted on the body times displacement of center of mass is a quantity of work done on the body—even though it might have been previously emphasized that a force undergoing no displacement is a zero-work force. The normal force in both these cases is certainly a zero-work force. It is incorrect and very misleading to imply that the wall or the ground do work on the body when the actual energy transformations are completely internal and no external work is being supplied. It is equally incorrect to suggest that the frictional force acting on the tires of an accelerating car does work on the car. If work were supplied from an external source in any of these cases, we would need neither food nor fuel.

In dealing with the energy transformations taking place when a box is being pushed along the floor at constant velocity against an equal and opposite frictional force, it is frequently implied that the frictional force times center of mass displacement of the box is a quantity of work taken out of the system (the box). If that were the case, the net work done on the box would be zero, and its rise in temperature would be a clear violation of energy conservation. Where does the increase in thermal internal energy come from?

The only correct and logical way of getting around these “infelicities” is to invent the concept of “internal energy” and introduce what amounts to the First Law of Thermodynamics right at the start. This can be done without excessive abstraction and without the more sophisticated mathematical apparatus usually associated with formal thermodynamics.

My intent in writing this monograph is to show how such an introduction of the energy concepts can be effected in an introductory physics course at calculus-based level. Rectilinear kinematics and dynamics are assumed as background. Although I started trying to treat energy by itself, I soon found that I could achieve satisfactory coherence and continuity only by treating conservation of momentum and mass as well as energy. Hence a monograph on the classical conservation laws. Readers will note that, being consistent with an educational position I have advocated throughout my entire academic career, I have included some modest measure of historical, philosophical, and epistemological background along with the conceptual development.

Some teachers may want to use this presentation as a text in itself, but that is certainly not necessary. It is quite understandable that many teachers will probably prefer to use a standard text that fits an entire course. They might, however, find it desirable to modify the text presentations in such a way as to obviate the incorrect and misleading portions. In some instances, this monograph might be useful as collateral reading. In any case, I would strongly recommend that it be read by the abler students so as to insure a firmer grasp of the energy concepts than they are otherwise likely to acquire.

Users should note that the questions and problems that are embedded in the text itself are not simply homework questions to be assigned or ignored at will. They are essential study guides that lead a student to stop, think, interpret, and digest the ideas being developed. They are indispensable to serious study and understanding. Without such practice, students do not construct the necessary concepts and insights and do not learn how to learn as mature adults. It has been shown repeatedly that conventional end of chapter problems are not sufficient to help develop such capacity for learning. Following the study guide supplied herein can, with the help of a teacher and cooperative group discussion, provide at least a start for more mature learning.

Contents of Part III

CHAPTER 1	LINEAR MOMENTUM	1
1.1	Introduction	1
1.2	Rectilinear Collisions of Two Bodies	2
1.3	Classification of Collisions	5
1.4	Viewing Perfectly Elastic Collisions From Different Frames of Reference	8
1.5	Viewing Perfectly Inelastic Collisions From Different Frames of Reference	12
1.6	Inductive Versus Deductive Reasoning and the Discovery of Natural Laws	14
1.7	Conservation of Momentum in Collisions	15
1.8	A Principle of Relativity	18
1.9	Collisions and Newton's Second Law	20
1.10	Impulse and Momentum Change for a Single Body Under More Than One Force	25
1.11	Impulse and Momentum Change for a Ball Striking a Wall	27
1.12	Impulse and Momentum in a Closed System	30
1.13	The Law of Conservation of Momentum	33
1.14	Center of Mass of a System of Particles	37
1.15	Questions and Problems	40
CHAPTER 2	INTERACTION, SYSTEM, STATE, AND CONSERVATION OF MASS	45
2.1	Introduction	45
2.2	System	47
2.3	State of a System and Changes in State	48
2.4	Pressure	51
2.5	Pressure in Fluids	54
2.6	Thermometers and Temperature	59
2.7	Thermal Equilibrium	60
2.8	The Concept of "Conservation"	62

2.9	Lavoisier and the Law of Conservation of Mass	63
2.10	Questions and Problems	65
 CHAPTER 3 THE CONCEPT OF "HEAT"		67
3.1	Distinction Between "Heat" and "Temperature"	67
3.2	Measuring Amounts of Heat Transferred	70
3.3	Specific Heat and Heat Capacity	72
3.4	Refinement of the Specific Heat Concept	75
3.5	Phase Change and Latent Heat	77
3.6	Heat is <i>Not</i> a Function of State	80
3.7	Temperature Changes Without Transfer of Heat	81
3.8	Questions and Problems	82
 CHAPTER 4 ENERGY		83
4.1	Change in the World Around Us	83
4.2	Review of Impulse, Momentum, and Collisions	87
4.3	Horizontal Acceleration of a Single Particle Under Constant Forces	89
4.4	Vertical Displacement of a Particle Under Constant Forces	94
4.5	Displacement of a Particle Under Action of a Varying Force	96
4.6	Displacement of a Particle Against the Restoring Force of a Spring	99
4.7	Vocabulary: Work and Kinetic Energy	102
4.8	Potential Energy	104
4.9	Units and Dimensions	109
4.10	Perfectly Inelastic Collisions	110
4.11	Using the New Vocabulary	111
4.12	Models for the Nature of Heat	114
4.13	Rumford's Attack on the Caloric Model	116
4.14	The Quantitative Relation Between Work and Heat	118
4.15	Joule's Enunciation of the Principle of Conservation of Energy	121
4.16	Extended Bodies and Systems as Opposed to Point Masses	123
4.17	Heat, Work, and Change of State: The First Law of Thermodynamics	124
4.18	The Varieties of Internal Energy	127
4.19	Dealing With Extended Systems	129
4.20	The Block and Spring Without Friction	130
4.21	Person Jumping Vertically Upward	133

4.22 Work and Heat in the Presence of Sliding Friction 137

4.23 Logical Status of the Conservation Laws144

4.24 Questions and Problems148

Index 151

Chapter 1

Linear Momentum

1.1 INTRODUCTION

Galileo, in the treatise *Discourses on Two New Sciences*, published in 1638 and written near the end of his life, put forth in clear mathematical form the science of “kinematics” essentially as we know it today and as you studied it in your initial study of physics.¹ This science, through generation of the concepts of velocity (including *instantaneous* velocity) and acceleration, provides a description of accelerated motion without inquiring into the causes of acceleration or into how one might predict what the accelerations would be in various circumstances. It was only with the addition of Newton’s laws of motion (and the associated concepts of force and mass) that it became possible to predict accelerations. This latter science is called “dynamics.”

In the *Discourses*, one of the participants raises the question as to what might be the causes of acceleration in free fall and in the motions of celestial bodies, but Galileo’s spokesman puts this discussion aside, saying “The present does not seem to be the proper time to investigate the cause of acceleration of natural motion [free fall] concerning which various opinions have been expressed by various philosophers. . . . At present it is the purpose of our Author merely to investigate and to demonstrate some of the properties of accelerated motion whatever the cause of this acceleration may be” Herein lies one aspect in which we have come to regard Galileo as a founder of modern science: He explicitly limits the scope of the inquiry to be conducted, deliberately defining a restricted problem (kinematics) through which progress can be made and insight acquired, rather than attacking too large and complex a problem (dynamics) that could not, at that point, be penetrated. Such self-conscious limitation of the scope of inquiry had not been invoked by Galileo’s predecessors. Stillman Drake (a Galileo expert and

¹Rather than using algebra and graphs in obtaining the kinematic relations, Galileo used the geometrical representations and proportional reasoning common in his own time. The techniques you employed to obtain the same results were invented somewhat later. The concepts and physical insights are, however, exactly the same.

translator of the *Discourses*) remarks that “Rejection of causal inquiries was Galileo’s most revolutionary proposal in physics, inasmuch as the traditional goal of that science was the determination of causes.”

As you continue your study of science, you will see, through one illustration after another, that there exists no one method or formula or prescription for an infallible procedure for scientific inquiry. But fruitful and successful steps in the winning of new knowledge sometimes have broad characteristics in common. One of the most clearly notable characteristics of modern scientific inquiry is the art (it is not itself a science) of limiting the scope of inquiry in such a way as to insure winning one significant step of understanding at a time, avoiding the distraction, confusion, and blocking effects of premature or irrelevant questions. This procedure, however, is not foolproof, and in some cases may serve to conceal important aspects and impede solution of a problem. Deciding when and to what extent to limit the scope of inquiry so as to gain a breakthrough is still the hallmark of individual genius.

Galileo was by no means insensitive to the importance of the question about causes of acceleration and was certainly not uninterested in the problem. He foresaw that progress in such inquiry might be made through the study of what was called “impact” or “percussion,” that is, collisions of bodies in which motions were altered, and he made some tentative stabs in this direction himself by studying the impact of falling water. The study of collisions was pursued throughout the 17th century by many investigators, the principal focus being on the simplest case—so-called “rectilinear collisions” in which spheres collided along their line of centers. Books of the time even show pictures that imply the firing of one cannon ball at another resting on a pedestal. Newton himself engaged in studies of collisions (using colliding pendulums) and appeals, at various points in the *Principia*, to then known empirical results.

Experimental observations showed that a key factor in predicting the outcome of a collision was the “quantity of motion” of each body participating in the interaction, “quantity of motion” being the name then given to the product mv of the mass and instantaneous velocity of an object. We now call this product the “momentum” of the object. It turns out that momentum plays a very deep role in dynamics, deeper than we were able to discern in our earlier study of Newton’s laws of motion confined to use of the second law in the form $\vec{F}_{\text{net}} = m\vec{a}$. It is this deeper role that we begin to study in this chapter.

1.2 RECTILINEAR COLLISIONS OF TWO BODIES

Consider two objects, which we shall treat as single particles, moving along their line of centers on a level surface as illustrated in Figs. 1.2.1.(a) and (b). (We wish, initially, to confine collisions to the line of centers so as to deal with the simplest possible case, that of motion in one, rather than two, dimensions. We call such collisions “rectilinear.”)

What changes in motion take place when the two particles collide? You can conduct simple experiments of this kind in a variety of ways: Billiard balls, large glass marbles, or steel ball bearings can be rolled so as to collide—preferably along a groove or track in order to insure rectilinear collision. Pendulum bobs of the same or different masses can be set up so as to collide along their line centers at the bottom of their swing. Laboratory carts (or gliders on an air track) can be equipped with spring bumpers (or with magnets mounted so as to repel each other). In the case of carts and gliders you can vary the masses of the colliding objects by loading or unloading them; you can also arrange to have them stick together and move off as one unit after collision instead of springing apart.

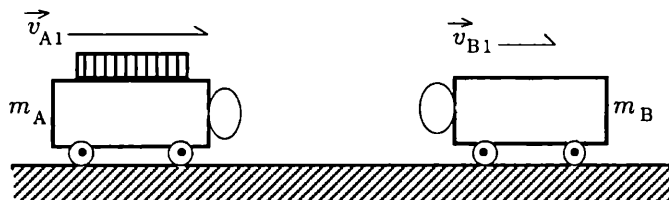


Figure 1.2.1(a) Carts A and B just before collision on a level table. The carts are equipped with very “soft,” flexible bumper springs.

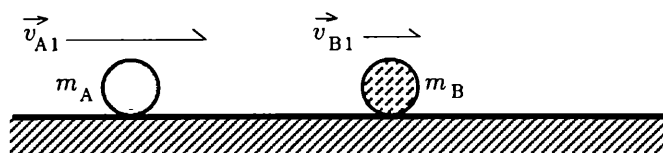


Figure 1.2.1(b) Steel spheres just before collision along their line of centers while rolling in a level groove.

Start in by making purely *qualitative* observations without worrying about numerical measurements. What experiments would you perform? How good are your powers of observation?

Following are some of the most basic qualitative regularities that you can discern in such simple experiments. (Velocities carry their usual algebraic signs with positive direction toward the right.)

- (a) With body B initially stationary ($\vec{v}_{B1} = 0$) and body A moving to the right at positive velocity \vec{v}_{A1} in our laboratory frame of reference:
 - (1) If $m_A = m_B$, body A stops ($\vec{v}_{A2} = 0$) and body B moves forward with velocity \vec{v}_{B2} very nearly equal to \vec{v}_{A1} .
 - (2) If $m_A > m_B$, both bodies move to the right after collision, B moving more rapidly than A and A moving less rapidly than it did *before* collision (i.e., both velocities positive and $\vec{v}_{B2} > \vec{v}_{A2}$).

- (3) If $m_A < m_B$, body A bounces back to the left while B moves off to the right (i.e., \vec{v}_{A2} becomes negative and \vec{v}_{B2} positive).
 - (4) If the bodies stick together on collision instead of bouncing apart, the combination moves to the right with a positive velocity \vec{v}_2 smaller than \vec{v}_{A1} regardless of the relative masses.
- (b) With bodies A and B initially moving toward each other with equal and opposite velocities in our laboratory frame of reference (i.e., \vec{v}_{A1} positive and \vec{v}_{B1} equal in magnitude but negative):
- (1) If $m_A = m_B$, and the bodies do *not* stick together, they bounce apart with equal and opposite velocities, i.e., they exchange their velocities so that $\vec{v}_{B2} = \vec{v}_{A1}$ (both positive) and $\vec{v}_{A2} = \vec{v}_{B1}$ (both negative).
 - (2) If $m_A = m_B$ and the bodies *do* stick together on collision, they both come to a dead stop, the final velocity of each one becoming zero (i.e., $\vec{v}_{A2} = \vec{v}_{B2} = 0$).

Question 1.2.1 Be sure to perform some qualitative experiments of this kind. Actually seeing the collisions and establishing the patterns that occur will greatly improve your intuitive feeling for the phenomena we shall be discussing in this chapter. Address the following questions as you make your observations:

- (a) Observe some cart (or air glider) collisions, such as those suggested in Fig. 1.2.1(a), in which $\vec{v}_{B1} = 0$ and in which the bumpers are so very “soft” that you can follow the deformation as they bend and then spring back. Sketch what the graph of force versus clock reading must look like for each body over the interval of contact. Sketch the corresponding graphs of acceleration and velocity. Then sketch similar graphs for the case of billiard balls or ball bearings, as suggested in Fig. 1.2.1(b), where it is impossible to see the actual deformations during the interval of contact. What defines the time interval of interaction in such collisions? That is, when does the interaction begin and when does it end? How do the time scales in the “soft” and “hard” collisions compare with each other?
- (b) What happens in the situation of Fig. 1.2.1(a) if you replace the springs by balls of putty or patches of velcro so that the carts stick together on collision instead of springing apart?
- (c) Now consider the collisions of carts or gliders equipped with magnets mounted so as to repel each other: What is happening in absence of physical contact? What do the various graphs [corresponding to those in part (a)] look like in such cases? What is the time interval of interaction, and how is this represented on the graphs? (Note that the objects need never touch each other, yet we still call this a collision.) What connection do you see between such collisions and what must happen when atoms or molecules of a gas collide with each other or with the atoms in the walls of their container?

- (d) Suppose the magnets are mounted so as to attract rather than repel each other (or imagine the carts to be carrying unlike electrical charges.) Describe what might happen in some possible collisions and sketch corresponding force versus clock reading graphs. How might you rig an analogous mechanical experiment (with springs or rubber bands)?
- (e) What influence and effects do you ascribe to friction in these various experiments?

Question 1.2.2 (a) Without trying to make specific or detailed predictions, visualize qualitatively the kinds of things that might happen when two bodies collide with each other (not necessarily along their line of centers) in three-dimensional space. Include in your visualizations the possibility of rotations and the rattling around of internal parts. (b) In the light of your visualizations in part (a), describe the kinds of things that might happen when atoms or molecules of a gas collide with each other.

1.3 CLASSIFICATION OF COLLISIONS

If you review the observations you have made of the very simplest cases of rectilinear collisions, you will see that the collisions can be separated into several distinct classes. (This is exactly what was done by the early observers who initiated such studies.) One clearly defined, extreme class is that of the collisions in which the colliding objects stick together and move off as a unit ($\vec{v}_{A2} = \vec{v}_{B2} = \vec{v}_2$) after collision. In such cases, we recognize that, whatever may be the velocity of A relative to B before collision ($\vec{v}_{A1} - \vec{v}_{B1}$), the velocity of A relative to B after collision ($\vec{v}_{A2} - \vec{v}_{B2}$) is zero. If body B is initially stationary ($\vec{v}_{B1} = 0$), the final velocity \vec{v}_2 of the combined pair is always smaller in magnitude than \vec{v}_{A1} . Such collisions are given the name “perfectly *inelastic*.” A second extreme class is that of collisions in which the velocity of A relative to B is reversed in direction but remains unchanged in magnitude; i.e.,

$$\vec{v}_{A2} - \vec{v}_{B2} = -(\vec{v}_{A1} - \vec{v}_{B1}) \quad (1.3.1)$$

This class of collisions is given the name “perfectly *elastic*.” Let us connect this classification with some of our earlier observations.

Question 1.3.1 In the algebraic forms in Eq. 1.3.1, positive direction is being taken toward the right, and individual velocities are in the laboratory frame of reference. Body B is being taken as the reference object, and the *relative* velocity is that of body A relative to body B. Verify this by examining the circumstances in which the velocity of A relative to B ($\vec{v}_{A1} - \vec{v}_{B1}$) would be positive and the circumstances in which it would be negative. Note, for example, that if both bodies are moving toward the right and $\vec{v}_{A1} - \vec{v}_{B1}$ is negative, the bodies cannot collide. What would the comparative sizes of \vec{v}_{A1} and \vec{v}_{B1} have to be

if a collision is to take place with both bodies moving toward the left? By illustrating with specific examples and diagrams, argue that, regardless of the direction in which the bodies are moving, collision can take place only if the quantity $\vec{v}_{A1} - \vec{v}_{B1}$ is positive and that collision cannot take place if the quantity is negative.

With simple collisions of identical objects ($m_A = m_B$), you may have observed the following special case: Body A, with initial velocity \vec{v}_{A1} , collides with stationary body B. Body A stops while body B moves off with a velocity equal to the initial velocity of A ($\vec{v}_{B2} = \vec{v}_{A1}$), i.e., the bodies exchange velocities relative to the laboratory.

Question 1.3.2 Argue that, in this case, the velocity of A relative to B changes direction but remains unchanged in magnitude. Show that the observed behavior is correctly expressed by Eq. 1.3.1.

You may also have observed the following special case: Identical bodies A and B approach each other with equal and opposite velocities of magnitude $|\vec{v}_1|$ relative to the laboratory. On collision, they bounce apart and recede from each other with equal and opposite velocities of magnitude $|\vec{v}_1|$ relative to the laboratory.

Question 1.3.3 Argue that, in this case, the velocity of A relative to B is unchanged in magnitude but reverses direction on collision. Show that Eq. 1.3.1 correctly describes the observed behavior in the collision.

If the colliding bodies do not have equal masses, it is more difficult to discern the connection between the initial and final conditions without making numerical measurements, but when such measurements are made, they show that, in the ideal limit, Eq. 1.3.1 continues to apply. As indicated above, collisions obeying this equation are classified as “perfectly elastic.” In actual fact, perfectly elastic collisions do not occur with the macroscopic objects we encounter in everyday experience (i.e., those consisting of huge numbers of atoms and molecules) although one comes quite close with objects made of materials such as steel or glass or ivory. (In the 17th and 18th centuries, collisions utilizing such materials and coming close to satisfying Eq. 1.3.1 were referred to as “hard collisions.”) Perfectly *inelastic* collisions are, however, a common occurrence.

Question 1.3.4 Cite some examples of perfectly inelastic collisions you have observed or can imagine in ordinary experience.

In light of the two extreme classes defined above, we can define a third class of collisions that might be called “partly elastic.” This class includes all the common occurrences in which the colliding objects do not stick together but

do not quite satisfy Eq. 1.3.1. (We shall see later, as we develop the energy concepts, that the terms “inelastic” and “elastic” have to do with the presence or absence of energy transformations involving increases in temperature. We need not, however, be concerned with this aspect at present.)

Although perfectly elastic collisions never occur in cases of mechanical contact between macroscopic objects, perfectly elastic collisions can occur in interactions between atoms, molecules, or subatomic particles on the microscopic scale. Such interactions can be perfectly elastic over certain ranges of intensity but are not always necessarily so.

Question 1.3.5 Referring back to your initial stab in Question 1.2.2, you might continue visualizing and discussing such situations without feeling that you must immediately perceive all the “right” answers. Don’t hesitate to start engaging in some speculation and conjecture: What would a perfectly elastic collision be like in such circumstances? What might be the nature of an inelastic collision; i.e., what might happen, or not happen, to internal structures or motions in an atom or molecule on collision with another?

In the light of the classifications we have defined above, it is apparent that most of the macroscopic contact collisions we observe in the world around us (those collisions in which the objects do not stick together) are to be classified as *partly inelastic*. Since the relative velocity of the colliding objects is unchanged in magnitude in perfectly elastic collisions and becomes zero in perfectly inelastic ones, it is clearly possible to invent a way of measuring the degree of elasticity (or inelasticity) of a collision by simply using the ratio of initial to final relative velocity magnitudes as the measure. This is precisely what is done, and the ratio is given the name “coefficient of restitution” and the symbol c . Thus the coefficient of restitution is defined by

$$c \equiv \frac{|\vec{v}_{A2} - \vec{v}_{B2}|}{|\vec{v}_{A1} - \vec{v}_{B1}|} \quad (1.3.2)$$

Question 1.3.6 Verify that Eq. 1.3.2 says that $c = 1$ for perfectly elastic collisions and that $c = 0$ for perfectly inelastic ones.

Question 1.3.7 A ball of putty is dropped from rest at a height of 2.0 m above the floor and strikes the floor without bouncing back. Use Eq. 1.3.2 to determine the coefficient of restitution for this collision.

Question 1.3.8 A ball is dropped from rest at a height of 2.0 m above the floor and rebounds to a height of 1.6 m. Calculate the coefficient of restitution for this bounce by first deriving a general *algebraic* expression for the coefficient of restitution when an object dropped from rest at a height H_1 rebounds to a height H_2 . (Note that you must go back to the relevant connection between velocity and vertical displacement in free fall.)

Question 1.3.9 Estimate the coefficient of restitution in some simple collision you actually set up (or at least visualize.)

1.4 VIEWING PERFECTLY ELASTIC COLLISIONS FROM DIFFERENT FRAMES OF REFERENCE

In building up his arguments concerning the superposition of horizontal and vertical velocities in projectile motion, Galileo repeatedly referred to observations one might make on a ship moving at uniform velocity: An object dropped from the top of a mast would appear, to observers on the ship, to fall straight down along the mast, while, to observers on the shore, it would appear to move in a parabolic path. The only difference, Galileo argued, lies in addition or subtraction of the uniform horizontal velocity of the ship and not in alteration of the physics of the phenomenon. In the light of Galileo's influence, it is not surprising that 17th century investigators of collisions proceeded to ask similar questions about the laws of impact: Is the dynamical phenomenon changed in any intrinsic way if a given collision experiment is performed on a uniformly moving ship? And experiments seem to have been performed to test the expectation that there would be no intrinsic change. The Dutch scientist Christian Huygens then proceeded to use, in a most significant way, the "thought experiment" of observing collisions both from the frame of reference of the moving ship and from the frame of reference of the shore. In doing so, he obtained a deep insight, but his method of analysis remained neglected until, at the end of the 19th century, the crisis that led to the birth of Einstein's theory of relativity led to new appreciation of his approach. We shall develop Huygens's analysis both for the sake of acquiring his insight and for the sake of preparing ourselves for the study of relativity. [Note to the student: It is virtually impossible to follow the analysis in the rest of this section and in the following section by passive reading of the text. If read in this fashion, it is likely to seem confusing and unintelligible. The best way to study this material is to follow it step by step with your own pencil and paper work, drawing the diagrams, and writing out, and even anticipating, each velocity relation as it comes along. In doing so, you will find the reasoning to be clear and simple, and you will also be acquiring practice that will prove very valuable in subsequent study of relativity.] We start, as Huygens did, with the plausible assertion that "equal, hard bodies" (meaning perfectly elastic bodies of equal mass), approaching each other with equal and opposite velocities rebound with velocities unchanged in magnitude as sketched in Fig. 1.4.1. This is fully consistent with our simple initial observations.

This starting point is, of course, arbitrary, but we do need a starting point, and this one is consistent with our deep sense of symmetry in natural phenomena and is at least qualitatively supported by observations such as those you made in Sect. 1.2. (It is this same sense of symmetry, for example, that leads us to expect identical objects, hanging at equal distances from the pivot point of a see saw, to balance each other. We would be very surprised by any other outcome and would be very certain that the two objects were not identical if they failed to balance.)

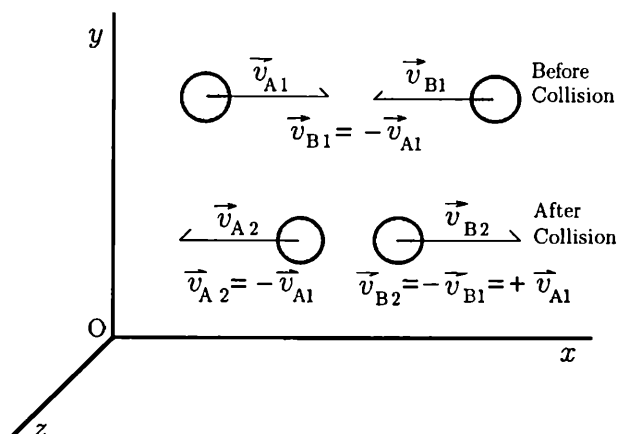


Figure 1.4.1 Perfectly elastic collision of bodies of equal mass approaching each other with equal and opposite velocities in reference frame O . Velocity \vec{v}_{A1} is taken as positive.

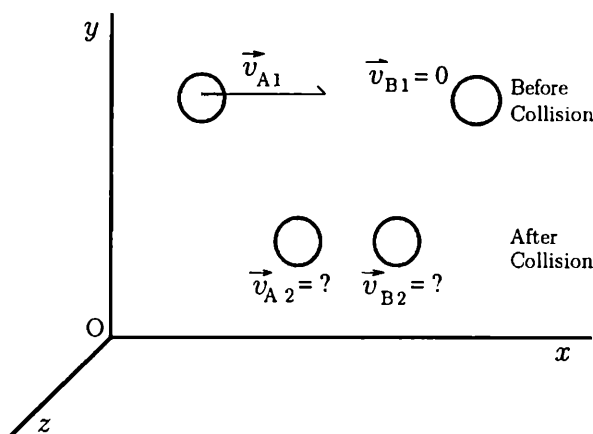


Figure 1.4.2 Perfectly elastic collision between identical bodies of equal mass. Body A with initial velocity \vec{v}_{A1} collides with initially stationary body B in reference frame O .

We next ask the question: What would we expect to happen if body B is initially stationary and an identical body A approaches with velocity \vec{v}_{A1} as in reference frame O in Fig. 1.4.2? We already know from our preliminary observations that, in elastic collisions under these circumstances, body A stops, and body B goes off with the initial velocity of body A ; i.e., the two bodies appear to exchange velocities. The question Huygens was raising is the following: Is there a simple connection between the collision represented in Fig. 1.4.1 and that in Fig. 1.4.2 or are the two circumstances independent and unconnected? In other words, given verification of the symmetrical and

expected behavior in the situation in Fig. 1.4.1, can we *predict* what we observe happening in the situation of Fig. 1.4.2? Huygens showed that not only are these situations deeply connected; they are essentially identical because they can be made identical by simply viewing the collision in Fig. 1.4.2 from another frame of reference.

The strategy is to adopt a new frame of reference O' in which the two bodies in Fig. 1.4.2 appear to approach each other with equal and opposite velocities. They must then bounce apart with equal and opposite velocities in this frame as they do in Fig. 1.4.1. The frame O' in which bodies A and B appear to move toward each other at equal and opposite velocities is one that moves to the right at velocity $\vec{v}_o = \vec{v}_{A1}/2$ relative to frame O. Let us examine this case with the help of Fig. 1.4.3.

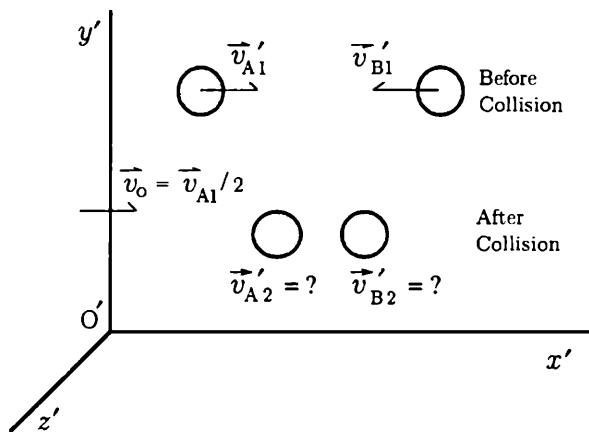


Figure 1.4.3 Frame O' moves to the right at velocity $\vec{v}_o = \vec{v}_{A1}/2$ relative to frame O of Fig. 1.4.2. Bodies A and B appear to approach each other in O' at equal and opposite velocities and bounce apart with equal and opposite velocities, all with magnitude $|\vec{v}_{A1}/2|$.

Since frame O' moves to the right at velocity \vec{v}_o relative to frame O, any velocity \vec{v}' in O' is related to any velocity \vec{v} in frame O by the relation $\vec{v}' = \vec{v} - \vec{v}_o$. Since \vec{v}_o is taken to be $+\vec{v}_{A1}/2$, the two bodies appear to approach each other in O' at equal and opposite velocities of magnitude $|\vec{v}_{A1}/2|$. (For example, $\vec{v}'_{A1} = \vec{v}_{A1} - \vec{v}_{A1}/2 = +\vec{v}_{A1}/2$, and $\vec{v}'_{B1} = 0 - \vec{v}_{A1}/2 = -\vec{v}_{A1}/2$.) In accordance with our assumption about what happens in Fig. 1.4.1, they must now bounce apart with velocities of magnitude $|\vec{v}_{A1}/2|$; i.e., $\vec{v}'_{A2} = -\vec{v}_{A1}/2$, and $\vec{v}'_{B2} = +\vec{v}_{A1}/2$. The final result is summarized in Fig. 1.4.4.

If we now view this bouncing apart from our original reference frame O, we see body A standing still since its velocity $\vec{v}_{A2} = \vec{v}'_{A2} + \vec{v}_o = -\vec{v}_{A1}/2 + \vec{v}_{A1}/2 = 0$. We would also see body B moving off to the right at velocity \vec{v}_{B2} since $\vec{v}_{B2} = \vec{v}'_{B2} + \vec{v}_o = +\vec{v}_{A1}/2 + \vec{v}_{A1}/2 = \vec{v}_{A1}$. This return to reference frame O is summarized in Fig. 1.4.5.

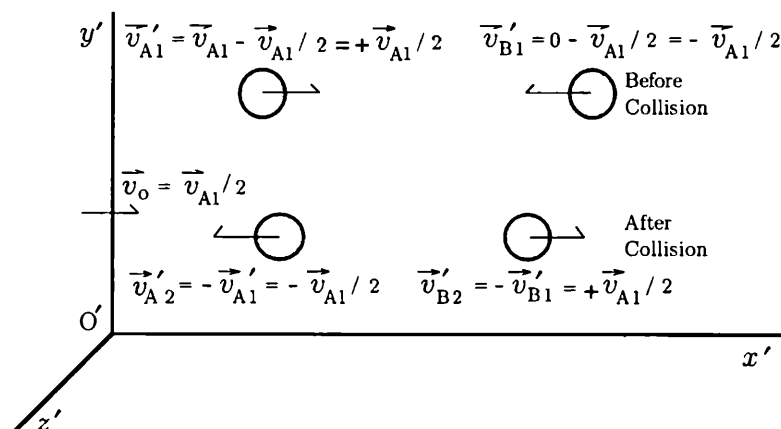


Figure 1.4.4 Collision in Fig. 1.4.2 in frame O' moving at velocity $\vec{v}_o = \vec{v}_{A1}/2$ relative to O .

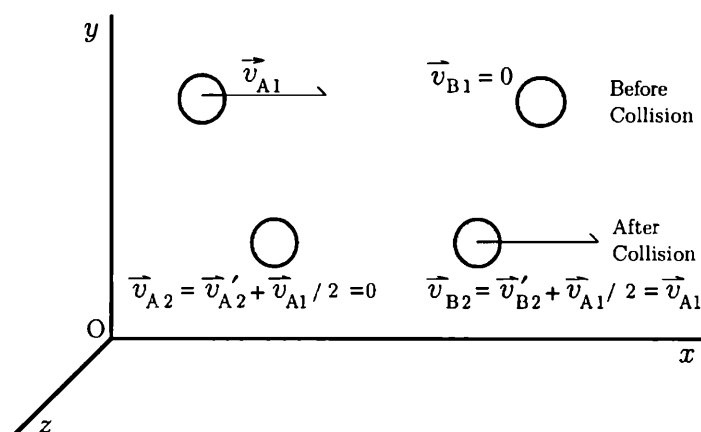


Figure 1.4.5 Collision in Fig. 1.4.4 in frame O' viewed from frame O .

Thus we find with Huygens that, if we take the symmetrical behavior observed in Fig. 1.4.1 to be fundamental, we can *predict* what will happen in the collision in Fig. 1.4.2 simply by following the argument involving changes in frames of reference. The two seemingly different collision events are *not* different and independent; they are essentially *identical* and merely *appear* different from different frames of reference. This implies a deep underlying regularity that might be expressible in mathematical form, that would have broad, general applicability, and that would eliminate the necessity of the special symmetrical starting points. In other words, it implies a law of nature that we should be able to guess and then test the validity of our guess by applying it to new cases.

1.5 VIEWING PERFECTLY INELASTIC COLLISIONS FROM DIFFERENT FRAMES OF REFERENCE

For the time being, let us continue by applying Huygens's method to a perfectly inelastic collision. We need to start with knowledge of a symmetrical case as a starting point, as we did with elastic collisions in the preceding section. We found that, when identical bodies approach each other with equal and opposite velocities and stick together on collision, the two objects come to a dead stop. This gives us our symmetrical starting point.

Now let us consider a collision, as in Fig. 1.5.1, in which body A, moving at initial velocity \vec{v}_{A1} in frame O, collides with stationary body B (i.e., $\vec{v}_{B1} = 0$), and the two bodies stick together. Our problem is to predict the final velocity \vec{v}_2 of the combined bodies based on our knowledge of the initial, symmetrical case.

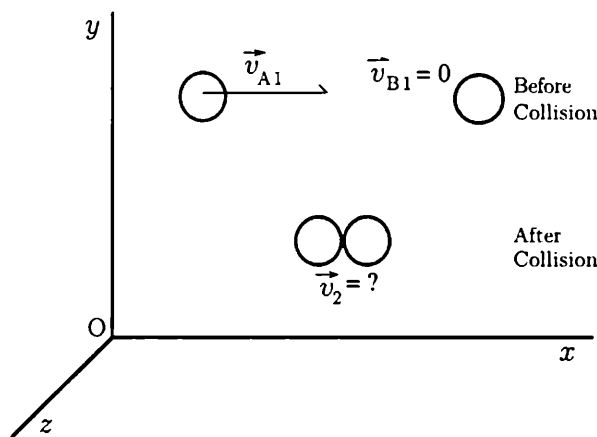


Figure 1.5.1 Body A moving at velocity \vec{v}_{A1} in frame O undergoes perfectly inelastic collision with identical stationary body B. What is the final velocity \vec{v}_2 of the combination?

Let us view the collision in Fig. 1.5.1 from a frame of reference O' moving relative to frame O at velocity $\vec{v}_o = \vec{v}_{A1}/2$ as in Fig. 1.5.2. The two bodies now appear to approach each other at equal and opposite velocities as in our examination in the preceding section, and, for this symmetrical case, we know that the final velocity \vec{v}'_2 of the combination is zero. Now going back from O' to O, we have the final velocity $\vec{v}_2 = \vec{v}'_2 + \vec{v}_o = 0 + \vec{v}_{A1}/2$. Thus we predict that, in the situation in Fig. 1.5.1, $\vec{v}_2 = \vec{v}_{A1}/2$, i.e., the two bodies, stuck together, will continue moving to the right after collision at half the initial velocity of body A. This is indeed what is observed to happen, and the prediction based on change of frame of reference relative to the symmetrical case is again successful.

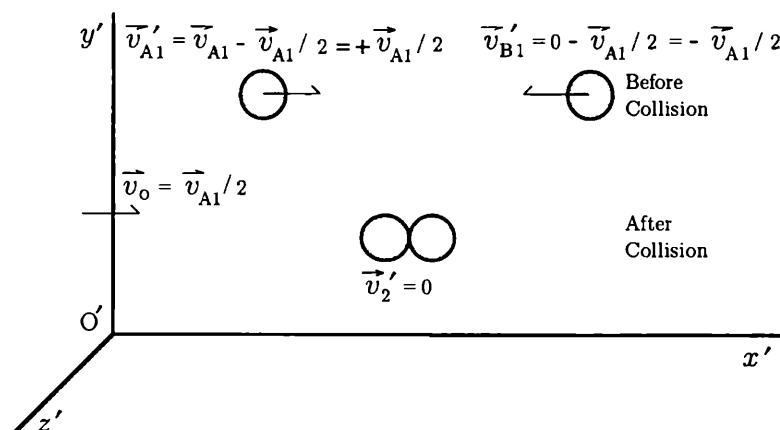


Figure 1.5.2 Perfectly inelastic collision in frame O (Fig. 1.5.1) viewed from frame O' moving at velocity $\vec{v}_0 = \vec{v}_{A1}/2$ relative to frame O .

Question 1.5.1 Apply Huygens's method to predicting what will happen in perfectly *elastic* collisions of bodies of equal mass moving in frame of reference O with velocities \vec{v}_{A1} and \vec{v}_{B1} respectively. (If \vec{v}_{A1} and \vec{v}_{B1} are both positive, \vec{v}_{A1} must, of course, be greater than \vec{v}_{B1} if the bodies are to collide.) Hint: Start by verifying the fact that the bodies will appear to be moving toward each other at equal and opposite velocities in a frame O' that is moving at velocity $\vec{v}_0 = (\vec{v}_{A1} + \vec{v}_{B1})/2$ relative to frame O ; i.e., at a velocity equal to the average of the velocities of the two bodies in frame O . (The answer in this situation turns out to be that $\vec{v}_{A2} = \vec{v}_{B1}$ and $\vec{v}_{B2} = \vec{v}_{A1}$; i.e., that the bodies exchange their velocities in this more general collision just as they do in the special cases previously considered!)

Question 1.5.2 Carry out a similar analysis for the situation described in Question 1.5.1 to predict the final velocity of the two bodies in a perfectly *inelastic* collision. (Note that you can check your result by seeing whether it reduces to the previously established value $\vec{v}_2 = \vec{v}_{A1}/2$ when $\vec{v}_{B1} = 0$.)

Question 1.5.3 Check the results associated with Fig. 1.4.1, Fig. 1.4.2, and Question 1.5.1 to verify the fact that in each perfectly *elastic* collision being considered, the velocity of body A *relative* to that of body B reverses direction in the collision but remains unchanged in magnitude. (Recall what was asserted in Sect. 1.3 about this regularity in perfectly elastic collisions.)

Question 1.5.4 Suppose we were to view the experiments discussed above from a frame of reference O' that was *accelerating* relative to O rather than moving at uniform velocity \vec{v}_0 . Would our method of analysis work? What difficulties would arise? Note how necessary it is to stay in an inertial frame of reference in order to exploit Huygens's approach!

Question 1.5.5 Show that all the results obtained for the cases considered in Question 1.5.3 and also for the perfectly inelastic collision in Figs. 1.5.1 and 1.5.2 are correctly described by the *algebraic* statement $\vec{v}_{A1} + \vec{v}_{B1} = \vec{v}_{A2} + \vec{v}_{B2}$. (By emphasizing the word *algebraic*, we are calling attention to the fact that the

equation being tested works *only* if the proper algebraic signs are used with the velocities; the equation is *not* valid for velocity *magnitudes*.)

1.6 INDUCTIVE VERSUS DEDUCTIVE REASONING AND THE DISCOVERY OF NATURAL LAWS

The laws of physics, which we pry out by study of natural phenomena, are not “derived” or “proved” as are theorems in algebra or geometry. They start as guesses or conjectures, based on sketchy or fragmentary evidence drawn from simple, special cases, and are then tested on as many other more complex and more general situations as possible to see whether they “work.” It is necessary to understand, however, that we are not talking about wild guesswork with no logical basis. Careful reasoning, including mathematical analysis, is an essential ingredient, but inferences drawn from the special cases may turn out to be of very little use through not being correct or applicable to any situations other than the ones from which they were obtained. In other words, the first step of logical reasoning in such instances is of the variety that we call “inductive”: we reason from the particular to the general, and, in such reasoning, one has no assurance that the results will be “correct” in the sense of having general validity.

Thus any generalization that is drawn inductively from the special cases must then be used as a tentative theorem or postulate from which one draws predictions about other, usually more general, situations that did not enter into the initial, sketchy evidence. We say that we “derive” results and predictions from the tentative generalization. Reasoning in this direction—from the general theorem to the particular case—is called “deductive.” This is the kind of reasoning you encountered in mathematics, where you started with axioms and postulates from which various theorems were deduced or “proved.”

This is actually an illustration of the meaning of the word “prove.” We “prove” some result by deductive reasoning from an accepted starting point. In other words, we show the result obtained to be a “true” consequence of the initial premises that were adopted. The initial premises, however (the axioms and postulates of Euclidean geometry, for example), cannot be proved or derived from some more fundamental starting point; they are arrived at inductively as reasonable starting points in their own right. Although Euclid’s postulates were originally considered unique, final, and inescapable, mathematical advances in the 19th century showed that other, less intuitively transparent, sets of postulates could lead to valid, but different, kinds of geometry that came to be called “non-Euclidean.”

Newton’s laws of motion and the law of universal gravitation, for example, were not derived from any initial, underlying premises or axioms. Newton arrived at them *inductively*, giving very little information about the steps of his own inductive reasoning. In the *Principia*, he then proceeds with volumes

of *deductive* reasoning, deriving all sorts of powerful results from the laws of motion and gravitation that agree with, and predict, observable phenomena. The fact that the scheme works is what ultimately justifies the initial inductive reasoning and leads to use of the word “law” in this context.

We outline the preceding thoughts about inductive and deductive reasoning because we are now about to take an inductive step. Out of the special, simple, restricted cases of collision examined in the preceding sections, we are going to join Huygens and his contemporaries in making an educated guess concerning the mathematical regularity underlying the phenomena. We shall then have to see whether the initial guess stands up under further deductive reasoning.

1.7 CONSERVATION OF MOMENTUM IN COLLISIONS

The collisions we examined in Sections 1.4 and 1.5 were extremely restricted in the sense that we kept using identical objects and did not investigate what happened in collisions of objects having different masses. Any generalization about mathematical order underlying collision phenomena must certainly take mass into account explicitly. How might we add this ingredient?

It is apparent to anyone who feels the impact of a moving object that the effect depends not only on the velocity but also on the mass of the object. The simplest arithmetical way of combining these two properties to obtain a larger number as either one increases is by means of the product mv . This number was invoked and used from the earliest days of the study of impact and was initially called “quantity of motion.” (This is how Newton refers to it in the *Principia*. Our modern name for this quantity is “momentum.”)

There is no assurance ahead of time that the number mv is the “right” one to use in extracting a generalization concerning collisions. Numbers such as m^2v , or m^2v^2 , or mv^3 , or \sqrt{mv} , or any other similar combination would also increase in size as m and v each increased. But, in science, we have learned always to start with the simplest possible forms because, over and over again, it has been found that nature seems to select simplicity underneath the apparent complexity. (We shall see in due course, however, that the combination mv^2 turns out to play its own important role in addition to that of the combination mv .)

Now for our inductive guess: We saw in Question 1.5.5 that all the rectilinear collisions between bodies of equal mass visualized in Sects. 1.4 and 1.5 obeyed the relation

$$\vec{v}_{A1} + \vec{v}_{B1} = \vec{v}_{A2} + \vec{v}_{B2} \quad (1.7.1)$$

(We use the vector notation to emphasize that this relation is valid *only* in the *vector* sense; i.e., the proper algebraic signs must be applied to the velocities in the rectilinear case, and the equation is *not* valid for velocity *magnitudes*.) Let us try the notion that this basic form might apply not just to velocity

alone but to *momentum* in the case of rectilinear collisions of bodies with different masses; i.e., let us guess (as did Huygens and his contemporaries) that all rectilinear collisions might obey the form

$$m_A \vec{v}_{A1} + m_B \vec{v}_{B1} = m_A \vec{v}_{A2} + m_B \vec{v}_{B2} \quad (1.7.2)$$

This equation has not been “derived” in any mathematical sense; it is an educated *guess* or *conjecture*, rooted in Eq. 1.7.1 and based on observations of the simple, very special cases of rectilinear collision of identical bodies. Whether or not it correctly describes the wider range of actual physical phenomena involving objects with different masses, however, can only be ascertained through further experiment and observation.

The fact is that all experiments with rectilinear collisions of objects with different masses turned out to support this conjecture, and it came to be regarded as a confirmed law or relation by the 1660s. It was even referred to as the basic “law of motion” until Newton changed the terminology by applying the term “laws of motion” to the three laws he enunciated in the *Principia*.

Let us note what Eq. 1.7.2 says: First it says that the concept that “works” and that we want to define by giving it the name “momentum” is not the magnitude mv but the vector quantity $m\vec{v}$ because it is the *vector*, not just the magnitude, that obeys the regularity that we have discovered. Secondly it says that, in rectilinear collisions between two objects, the total vector momentum of the two objects *before* collision is equal to the total vector momentum of the same two objects *after* collision, i.e., the total momentum of the combination of the interacting objects is preserved (does not change) despite the various individual velocity changes that take place. In modern terminology, we translate Eq. 1.7.2 as saying that “momentum is *conserved* in rectilinear collisions.”

In science, we say that a quantity is “conserved” when its numerical value is the same before and after changes that take place in a system being described. For example, because of the *vector* relation with which we are confronted, the total momentum of a system consisting of two objects of equal mass approaching each other with equal and opposite velocities is *zero*. In a perfectly elastic collision, the objects bounce apart with equal and opposite velocities, and the total momentum of the system is still zero after the collision. If the collision is perfectly inelastic, both objects stick together and come to a dead stop; the total momentum is still zero both before and after the interaction.

Of course, Eq. 1.7.2 holds only if the colliding bodies interact solely with each other and are not acted upon by external forces. In the experiments we have been describing and visualizing, friction acting on either or both objects would be an external force and would keep decreasing the magnitudes of all the individual momenta involved. Performing such collision experiments on a sloping plane would introduce as an external force the component of gravitational pull along the plane, and momentum would not be conserved in the

sense of Eq. 1.7.2. Similarly the equation would not hold if we ourselves pushed or pulled on either body during the experiment. If the bodies interacted with each other by carrying magnets or electric charges, the presence of a nearby magnet or nearby electric charge would introduce an external force on the colliding bodies, and Eq. 1.7.2 would not hold for the observed momenta. If we added or took away material from either body during the experiment (for example if the bodies were carts on which we could drop sand during the interaction), Eq. 1.7.2 would not hold.

We shall use term “system” to describe two or more objects that interact with each other to produce changes in motion (or changes in other properties or conditions of the bodies.) When no *external* force is acting and when no matter is going in or out of the system, we say that the system is “closed.” In the opposite situation, we say that the system is “open.” The assertion that momentum is conserved is valid only for a closed system.

Question 1.7.1 Let us visualize some simple changes taking place in everyday experience and examine them from the standpoint of conservation or non-conservation of certain familiar properties.

- (a) Suppose we take a lump of clay (or our childhood play material plasticene) and deform it into various shapes such as a round ball, a flat pancake, a cup, a long cylinder, without adding or subtracting any material (i.e., our system consisting of the lump of clay is closed). Is the volume of material conserved under these various changes? Is the exposed surface area conserved? Is the mass of material conserved? Explain your reasoning.
- (b) Answer the same set of questions for the case in which we start with a solid object such as a stone or a piece of glass and break it up into smaller pieces without adding or losing any material.
- (c) Answer the same set of questions for the case in which we start with a given amount of liquid (say water) and pour it into containers of different shapes.
- (d) Answer the same set of questions for the case in which a lump of sugar dissolves in water.
- (e) Answer the same set of questions for a case in which a sample of air is compressed in a bicycle pump.
- (f) Answer the same set of questions for a case in which a sample of liquid or solid is subjected to a fairly large change in temperature or pressure.
- (g) Answer the same set of questions for the case in which a piece of wood is burned.

Question 1.7.2 Suppose we have a system consisting of two bodies that undergo very nearly perfectly elastic rectilinear collision under ordinary circumstances. We now place a percussion cap on one of the bodies so that the cap goes off as the bodies collide.

- (a) Is this system “closed” in the sense defined earlier?

- (b) Is the collision “elastic” in the sense defined earlier? Explain your answers in some detail. [Answers: (a) yes, (b) no.]

What use can we make at this point of Eq. 1.7.2; i.e., what can we *predict* by making use of it? Normally, we would know the initial conditions in rectilinear collision (the masses m_A and m_B of the two bodies and their velocities \vec{v}_{A1} and \vec{v}_{B1}). We would like to predict the final two velocities, but that means that there are two unknowns, and we have only one governing equation. We need an additional restriction. In the special case of the perfectly inelastic collision we do have such a restriction since the bodies stick together:

$$\vec{v}_{A2} = \vec{v}_{B2} = \vec{v}_2 \quad (1.7.3)$$

Combining Eqs. 1.7.2 and 1.7.3, we have, for the special case of the perfectly inelastic collision:

$$m_A \vec{v}_{A1} + m_B \vec{v}_{B1} = (m_A + m_B) \vec{v}_2 \quad (1.7.4)$$

With Eq. 1.7.4 we can predict the outcome of any perfectly inelastic rectilinear collision since we have only one unknown, namely the final velocity \vec{v}_2 .

Question 1.7.3 Interpret Eq. 1.7.4 carefully in words:

- (a) What happens if the initial velocities are both positive and body A overtakes body B? Examine this situation in terms of what happens if the masses of the bodies are equal; if m_A is very much larger than m_B ; if m_A is very much smaller than m_B .
- (b) What happens if the velocity of B is negative while that of A is positive? In this connection, examine the predictions for various relative masses as in part (a).

It is evident from Question 1.7.3 that we have obtained quite a complete story for perfectly *inelastic* rectilinear collisions, but we remain stuck with two unknowns and only one equation for perfectly and partially *elastic* collisions. We must defer the completion of this part of the problem until we learn still more about the connection among initial and final velocities (or more about the role, hinted at earlier, of the quantity mv^2).

1.8 A PRINCIPLE OF RELATIVITY

Let us return for a moment to the analytical techniques we used in Sects. 1.4 and 1.5—namely Huygens’s method of viewing a given collision from two different frames of reference. What assumptions were we making in adopting this method? This is something we glossed over without detailed examination in our first approach, and we now better have a closer look. We can identify three especially significant assumptions:

- (1) We assumed that a particular physical law was being obeyed regardless of our state of uniform motion in a straight line relative to the events being observed. Specifically, we started with the observed *facts* that (a) in a perfectly elastic collision, bodies of equal mass, approaching each other with equal and opposite velocities, bounce apart with their initial velocities reversed and (b) that in a perfectly inelastic collision, they stop dead. We then *assumed* that these statements of regularity would be valid for collision events in *any* frame of reference moving at uniform velocity in a straight line, i.e., an inertial frame, one in which Newton's first law is known to hold without invention of fictitious forces. We cannot prove this assumption to be valid on the basis of some more fundamental principle. We are reasoning inductively, and we make the guess and hope that nature might either validate it or show us to be wrong.
- (2) We assumed that velocities \vec{v}' observed in frame O' are related to velocities \vec{v} in frame O by the simple arithmetic of $\vec{v}' = \vec{v} - \vec{v}_0$, where \vec{v}_0 is the velocity of frame O' relative to frame O . (This seems to make perfectly good physical sense in terms of our intuitive feelings about such velocities, but it in fact turns out to be invalid for situations that obtain at velocities approaching the velocity of light. We shall have to renounce this assumption when we begin studying what are called "relativistic phenomena" even though this assumption is perfectly valid for the cases in which we are currently interested.)
- (3) We assumed that the proper expression for the quantity involving both mass and velocity of a given body is the simple form $m\vec{v}$ regardless of the magnitude of the velocity in our frame of reference. (This also turns out to be an invalid assumption when we are concerned with objects moving at velocities approaching that of light although it is perfectly valid for cases of current interest.)

Assumption (1) is an example of a "principle of relativity." It asserts that a certain physical law or regularity (in this case a law of impact) remains "invariant" (unaltered) when viewed from a different inertial frame of reference. In the special cases examined in Sects. 1.4 and 1.5, it is readily apparent that the *numerical values* of velocity and momentum of bodies in a given collision are quite different when viewed from frames of reference O and O' ; momentum itself is therefore *not* invariant to change of frame of reference. In each of the illustrations, however, *total* momentum of the system before collision is equal to *total* momentum after collision as far as each frame of reference is concerned, even though the numerical value of the total momentum is quite different in each frame. What is invariant with respect to change of frame of reference is the fact of *conservation* of the property in an interaction rather than persistence of its numerical value for a given body or for the interacting

system. It is the *conservation* of momentum in an interacting system that obeys the principle of relativity and is invariant under change of inertial frame of reference while the numerical value of the momentum is not itself invariant.

Discussions of principles of relativity played a crucial role in physics toward the end of the 19th century. The problem was to identify precisely what laws, regularities, and properties were invariant to transformation from one inertial frame to another. We have just seen that momentum conservation seems to be an invariant. In so-called “classical” physics (to distinguish it from relativistic physics involving very high velocities), accelerations and masses of objects are both invariant to change of inertial frame of reference. Electrical charge turns out to be an invariant both classically and relativistically. Energy conservation turns out to be an invariant. Electromagnetic phenomena did not seem to fit consistently into the picture along with mechanical phenomena, and Einstein showed that the difficulties and contradictions could be eliminated by giving a dominant role to a principle of relativity in an extended and generalized form at the cost of abandoning the attractively simple arithmetic of assumptions (2) and (3). Through understanding of what we have done in the preceding sections, you will eventually be better prepared to understand the changes Einstein introduced.

Question 1.8.1 Develop an analysis to show that, while the instantaneous velocity \vec{v} of a body is *not* invariant to shifting from one inertial frame of reference to another, the acceleration \vec{a} of the body *is* invariant to such transformation. [Hint: go back to the basic *definitions* of velocity and acceleration and examine what happens under change of frame of reference from O to O'. [Hint: The statement $\vec{v}' = \vec{v} - \vec{v}_0$ says, immediately, that the velocity is not invariant. How is acceleration connected to the velocity? Does \vec{v}_0 enter into the relation between \vec{a} and \vec{a}' ?]

1.9 COLLISIONS AND NEWTON'S SECOND LAW

In the rectilinear collisions we have been discussing, each individual body undergoes a change in momentum—which means a change in velocity and therefore an acceleration. This, in turn, means that each body was subjected to a net force during the interval of interaction. Thus we have the possibility of examining momentum changes in terms of the connection between force and acceleration. By linking what we have learned in the preceding sections about the role of momentum in collisions to the dynamical theory studied earlier, we can develop better understanding of the new, and still relatively unfamiliar, concept of momentum.

Fig. 1.9.1 shows the horizontal forces acting on each of the carts in Fig. 1.2.1(a) during the interval of contact between the springs. (The vertical forces acting on the carts are not shown in order not to clutter up the diagram, and frictional forces are assumed to be negligible.) Newton's third law says that

$\vec{F}_{AB} = -\vec{F}_{BA}$. The magnitude of the force \vec{F} varies with time during the interaction; it starts at a value of zero at the instant of initial contact t_1 , rises to a maximum value at closest approach of the two carts and maximum compression of the springs, and drops back to zero at the instant t_2 of breaking of contact. To emphasize the fact that the force varies with time, we shall use the functional notation $\vec{F}(t)$. A sketch of a possible force-time history for body B is shown in Fig. 1.9.2. The time interval of contact between the bodies is $t_2 - t_1$.

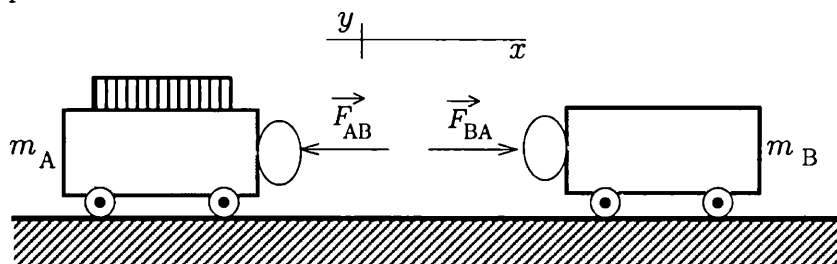


Figure 1.9.1 Horizontal forces acting on each of the two carts during the interval of interaction in the rectilinear collision in Fig. 1.2.1(a). Friction is assumed negligible. By Newton's third law $\vec{F}_{AB} = -\vec{F}_{BA}$.

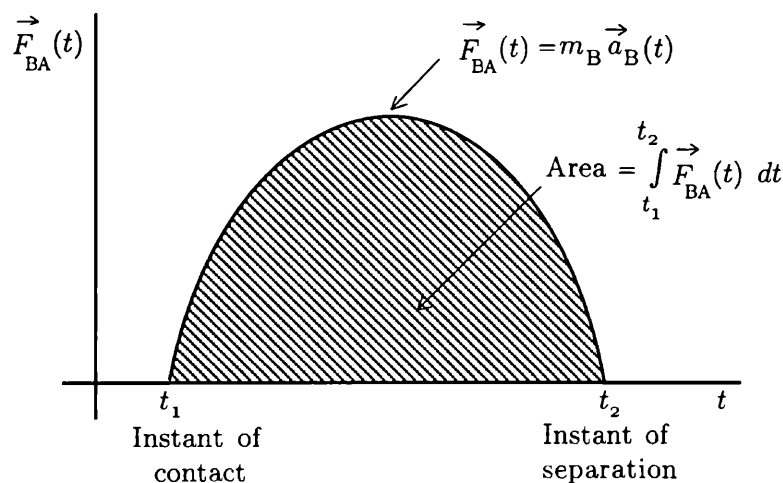


Figure 1.9.2 Time history of the variation of the magnitude of the horizontal force \vec{F}_{BA} acting on cart B during the interval of collision $t_2 - t_1$ in Fig. 1.9.1.

Question 1.9.1 Examine the shape of the graph sketched in Fig. 1.9.2. Does the algebraic sign assigned to the force make sense? The force is shown to be changing most rapidly at initial contact and having zero rate of change at its greatest value. Does this make physical sense? Why or why not? Does the symmetry sketched make physical sense? Why or why not?

Applying Newton's second law $\vec{F}_{x \text{ net}} = m\vec{a}_x$ to body B, we have

$$\vec{F}_{BA}(t) = m_B \vec{a}_{xB} \quad (1.9.1)$$

The horizontal force $\vec{F}_{BA}(t)$ varies in magnitude from instant to instant, and we do not know the functional relation. Eq. 1.9.1 therefore tells us very little about what happens to the velocity and momentum of cart B over the time interval of contact since it only contains the instantaneous acceleration, which also varies from instant to instant with the force. If we integrate varying acceleration, however, over any given time interval, we obtain the velocity change occurring in that interval.

Thus our knowledge of the relation between velocity and acceleration suggests that we should integrate both sides of Eq. 1.9.1 with respect to time over the interval between instants t_1 and t_2 because we will then obtain, on the right-hand side, the velocity change (and therefore the momentum change) of body B over this time interval. In this way we shall obtain information about momentum change from an analysis that began with Newton's second law of motion. As we perform the integration, let us recognize that the integral on the left-hand side will be equal to the shaded area under the force versus clock reading graph in Fig. 1.9.2. Carrying out the integration, and emphasizing the fact that we are dealing with accelerations and velocities in the x direction, we have

$$\int_{t_1}^{t_2} \vec{F}_{BA}(t) dt = \int_{t_1}^{t_2} m_B \vec{a}_{xB}(t) dt = m_B \vec{v}_{xB}(t_2) - m_B \vec{v}_{xB}(t_1) \quad (1.9.2)$$

Since the two velocities on the right-hand side of Eq. 1.9.2 are simply the velocities of body B before and after collision respectively, we can use the Δ notation and write Eq. 1.9.2 as

$$\int_{t_1}^{t_2} \vec{F}_{BA}(t) dt = m_B \vec{v}_{xB2} - m_B \vec{v}_{xB1} = \Delta(m_B \vec{v}_{xB}) \quad (1.9.3)$$

We should immediately recognize the right-hand side of Eqs. 1.9.2 and 1.9.3 as the change in the horizontal momentum of body B! Furthermore, the integral on the left-hand side represents, as we said earlier, the area under the force versus clock reading graph in Fig. 1.9.2.

We now introduce a new name, the meaning of which must be memorized: We give the name "impulse delivered by the given force" to the *area* under the force-time history for *any* force. To make such a calculation of area, we must know the history of variation, in other words we must have the entire graph *instant by instant*. We call such a calculation "path dependent." In the special case of a constant force \vec{F} acting for a time interval Δt , the impulse delivered by the force would simply be $\vec{F}\Delta t$, the word "constant" giving us the history that we must know in order to make the calculation of the area, which, in this special case, is simply a rectangle in the \vec{F} versus t coordinates.

Using our new terminology, we translate Eq. 1.9.3 as saying “the horizontal impulse delivered to body B by the single horizontal *net* force $\vec{F}_{BA}(t)$ is equal to the change in horizontal momentum of body B.”

Eq. 1.9.3 is a very remarkable equation in the following sense: In order to calculate the impulse delivered (the number on the left-hand side) we would have to know the *entire* history of variation of the force $\vec{F}_{BA}(t)$ *instant by instant*. The calculation is path dependent. But to calculate the number on the right hand side, we would only have to know two values: The final and the initial velocities of the body, and we need know nothing of the variation of velocity in between. The calculation on the right-hand side is *path independent*! Thus if, for example, we measured the velocities before and after collision and calculated the momentum change of body B, the number in our possession would be equal to the impulse (the area under the force-time graph), and we would know something about the force history even if we did not know the history instant by instant. (When you eventually study thermodynamics, you will encounter other situations in which a path dependent number on one side of an equation turns out to be equal to a path independent number on the other side. Such relations turn out to play a key role in calculation and prediction of energy transformations.)

A word about units: Since, in the SI system, mass is expressed in kg and velocity in m/s, the units of momentum must be kg m/s. There is no special name for this combination. Since force is expressed in newtons (N), the SI units of impulse must be N·s.

Question 1.9.2 Show, from basic definitions, that kg m/s and N·s are completely equivalent sets of units.

As indicated above, Eq. 1.9.3 says, in our new terminology, that the impulse delivered to body B by the single force acting on the body is equal to the change of momentum imparted to body B. We shall see shortly, however, that, if several forces act on an object simultaneously, each force delivers an impulse to the object but only the *net* impulse, i.e., that delivered by the *net* force, is equal to the change in momentum.

Note that, in the collision in Fig. 1.9.1, the force $\vec{F}_{AB}(t)$ acting on body A is, by Newton's third law, equal and opposite, instant by instant, to the force acting on body B, and the time interval of interaction is the same for both bodies. Since the force $\vec{F}_{AB}(t)$ is negatively directed, the force-time graph for body A is simply Fig. 1.9.2 flipped down over the x -axis, and the impulse delivered to body A must have the same magnitude as the impulse delivered to body B but the opposite algebraic sign. Thus the momentum change of body A must also be equal and opposite to that of body B, and the momentum change of A is therefore negatively directed. (Remember that impulse and momentum are both vector quantities.)

Question 1.9.3 Using the statements made in the preceding paragraph, translate them into symbols, and show:

- (a) That the statement about equal and opposite impulses is consistent with the restriction that the system, comprised of the two bodies in Fig. 1.9.1, is closed.
- (b) That the statement about equal and opposite momentum changes of the two bodies implies that the initial total momentum of the system is equal to the final total momentum of the system. Then compare all this with the story told in Section 1.7 and comment on the connections to be drawn between Sects. 1.7 and 1.9.

Suppose that we observe a collision such as that in Fig. 1.9.1 and measure the initial and final velocities of body B of known mass m_B . We therefore know the change in momentum $\Delta(m_B \vec{v}_{x B})$. Suppose also that this is a “slow” collision resulting from the presence of “soft” springs, and we are able to estimate the time interval of contact $t_2 - t_1$. Given the change in momentum $\Delta(m_B \vec{v}_{x B})$, we know the impulse that must have been delivered, but we do not know the time history. Let us ask ourselves the following question, however: What *constant* force acting for the same interval $t_2 - t_1$ would have imparted the same impulse?

Question 1.9.4 Translate the last question in the following way: Show that we are in effect asking “what height of rectangle has the same base and the same area as that of the actual force-time graph?” On a copy of Fig. 1.9.2, draw the rectangle whose height has the same area as the shaded area under the graph.

The constant force that imparts the same impulse in the same time interval (i.e., the height of the equivalent rectangle) is called the “time average force.” Using the bar over the symbol to denote “average,” we have

$$\bar{F}_{BA}(t) \equiv \frac{\text{Area}}{t_2 - t_1} = \frac{\int_{t_1}^{t_2} m_B \vec{a}_{x B}(t) dt}{t_2 - t_1} = \frac{\Delta(m_B \vec{v}_{x B})}{\Delta t} \quad (1.9.4)$$

or

$$\bar{F}_{BA} \equiv \frac{\Delta(m_B \vec{v}_{x B})}{\Delta t} \quad (1.9.5)$$

where the symbol \equiv should be read “is defined as.”

For example, if the change of momentum of body B on collision is found to have been $+12.6 \text{ kg m/s}$, and the time interval of contact is estimated to have been 0.20 s , the average net force acting on body B during the given interval must have been $+12.6 \text{ kg m/s} \div 0.20 \text{ s} = +63 \text{ N}$. The *net* impulse delivered to body B must have been $+12.6 \text{ N}\cdot\text{s}$. In this case the change of momentum of body A must have been -12.6 kg m/sec ; the average force acting on body A must have been -63 N ; the net impulse delivered to body A must have been $-12.6 \text{ N}\cdot\text{s}$.

Question 1.9.5 In the preceding numerical example let us consider the system comprised of the *two* bodies A and B. What has been the momentum change of the *system*? What was the net impulse delivered to the system?

Question 1.9.6 Suppose a net impulse of $18.6 \text{ N}\cdot\text{s}$, in the positive x -direction, is delivered to a 75 kg body in an interval of 0.15 s . At the instant the force is applied the body has a velocity of $+0.53 \text{ m/s}$ in the x -direction. Explaining your reasoning as you proceed:

- Calculate the average force exerted on the body.
- Calculate the momentum *change* imparted to the body.
- Calculate the final velocity of the body.
- What would have been the final velocity of the body if the initial velocity had been -0.53 m/s instead of $+0.53 \text{ m/s}$?

1.10 IMPULSE AND MOMENTUM CHANGE FOR A SINGLE BODY UNDER MORE THAN ONE FORCE

Having investigated the connection between $\vec{F}_{\text{net}} = m\vec{a}$ and change of momentum of a single body of mass m for the very simple special case in which only a single force acts to change the momentum of the body, we can now easily put together a somewhat more general statement. Let us consider the case in which two non-collinear forces \vec{F}_j and \vec{F}_k act on a body of mass m in the x - y plane as shown in Fig. 1.10.1.

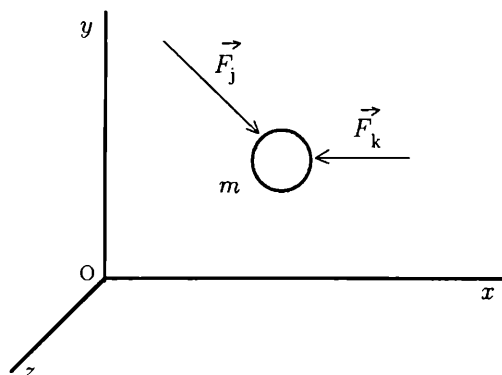


Figure 1.10.1 Body of mass m acted upon by two non-collinear forces \vec{F}_j and \vec{F}_k in the x - y plane.

We suppose the forces to act over the time interval between clock readings t_1 and t_2 . Let us resolve the forces \vec{F}_j and \vec{F}_k into their x and y components respectively and apply exactly the same analysis we carried out in connection with Eqs. 1.9.2 and 1.9.3. We obtain the following two equations:

$$\int_{t_1}^{t_2} \vec{F}_{x \text{ net}}(t) dt = m\vec{v}_{x2} - m\vec{v}_{x1} = \Delta(m\vec{v}_x) \quad (1.10.1)$$

$$\int_{t_1}^{t_2} \vec{F}_{y \text{ net}}(t) dt = m\vec{v}_{y2} - m\vec{v}_{y1} = \Delta(m\vec{v}_y) \quad (1.10.2)$$

where, in the particular case in Fig. 1.10.1, $\vec{F}_{y\text{ net}}$ would simply be equal to the y component \vec{F}_{yj} of force \vec{F}_j , and $\vec{F}_{x\text{ net}}$ would be equal to $\vec{F}_{xj} + \vec{F}_{xk}$, the algebraic sum of the x -components of \vec{F}_j and \vec{F}_k . Note that in the special case in this figure, \vec{F}_{yj} is a negative quantity, \vec{F}_{xj} is a positive quantity, and \vec{F}_{xk} is a negative quantity since the force components are vector quantities with appropriate algebraic signs. Therefore the two integrals on the left-hand sides of Eqs. 1.10.1 and 1.10.2 are also the x and y components of a vector quantity—the vector quantity to which we gave the name “impulse” in the preceding section.

Question 1.10.1 For the particular case illustrated in Fig. 1.10.1, the directions of the forces \vec{F}_j and \vec{F}_k have been specified, and hence the directions of the components are also specified. Suppose we use the symbol $|\vec{F}_x|$ to denote the *magnitude* of an x component of force and the symbol $|\vec{F}_y|$ to denote the magnitude of a y component. In the light of these definitions, argue that, for the particular situation defined in Fig. 1.10.1, $|\vec{F}_{y\text{ net}}| = -|\vec{F}_{yj}|$ and $|\vec{F}_{x\text{ net}}| = |\vec{F}_{xj}| - |\vec{F}_{xk}|$. How do these statements compare with those made in the preceding paragraph? Do they agree or disagree on algebraic signs? Explain your reasoning.

Since m is a scalar quantity and since the only other term in the momentum expression is the vector quantity \vec{v} , the momenta $m\vec{v}_x$ and $m\vec{v}_y$ must obey the same addition rules as do velocity components. Hence $m\vec{v}_x$ and $m\vec{v}_y$ are the x and y components of a total momentum vector $m\vec{v}$. The quantities $\Delta(m\vec{v}_x)$ and $\Delta(m\vec{v}_y)$ are therefore to be described as the “changes in the x and y components of the momentum vector $m\vec{v}$.”

It is necessary to learn to use our new vocabulary accurately and precisely if we are going to understand textbooks and problems, as well as each other. New vocabulary must be *memorized*; it can rarely be figured out. Following are a few examples of appropriate use of the new vocabulary:

- (a) We call the statements of Eqs. 1.10.1 and 1.10.2 connecting *net* impulse delivered to a body with the resulting momentum change of the body “the impulse-momentum theorem.”
- (b) In Fig. 1.10.1 we are dealing with external forces acting on a single body and not with a closed system. Hence momentum is *not* conserved in this situation. The momentum of the body keeps on changing as the forces act to deliver a net impulse.
- (c) We translate Eqs. 1.10.1 and 1.10.2 into words, saying that the *net* impulse delivered to a body by the combined forces acting on it is equal to the change of momentum of the body. Only a *net* impulse is equal to a change in momentum, however. Any individual force (or force component) has a time integral of its own, and this integral is called the “impulse delivered by that force,” but each such separate impulse is not necessarily equal to a momentum change of the accelerated body.

As a specific illustration of how to use the vocabulary, let us return to Fig. 1.10.1: The force component \vec{F}_{yj} acts alone in the negative y direction. It delivers a negative *net* impulse to the body, and this negative net impulse is equal to the component of change of momentum of the body in the y direction. The force component \vec{F}_{xj} , however, delivers an impulse to the body in the positive direction, but this impulse is *not* equal to the change of momentum of the body in the x direction. The force component \vec{F}_{xk} simultaneously delivers an impulse in the negative x direction, and the momentum change of the body in this direction is equal to the algebraic sum of the two separate impulses. The terminology is illustrated in the following equation for effects in the x direction in Fig. 1.10.1:

$$\int_{t_1}^{t_2} \vec{F}_{x\text{ net}}(t)dt = + \int_{t_1}^{t_2} |\vec{F}_{xj}(t)|dt - \int_{t_1}^{t_2} |\vec{F}_{xk}(t)|dt = m\vec{v}_{x\,2} - m\vec{v}_{x\,1} = \Delta(m\vec{v}_x)$$

Net impulse delivered by force components in x direction.	Impulse delivered by force component \vec{F}_{xj} .	Impulse delivered by force component \vec{F}_{xk} .	Change in x component of momentum of the body.
-------------------------------------------------------------	-------------------------------------------------------	-------------------------------------------------------	--------------------------------------------------

- (d) Finally, we note that the impulse-momentum theorem says that a given *net* impulse imparts the same momentum change to *any* body regardless of its mass. The velocity change would be smaller for larger masses, but the *momentum* change would be exactly the same!

It is interesting to note, in passing, that, in the *Principia*, Newton himself never put forth the second law of motion in the form familiar to us, namely $\vec{F}_{\text{net}} = m\vec{a}$. This form was introduced in the 18th century, after Newton’s death, through the work of the great mathematician Leonhard Euler. Newton worked with the form to which we have given the name “impulse-momentum theorem,” i.e., Eqs. 1.10.1 and 1.10.2, although he used a different system of symbols. When Newton uses the term “motive force” in the *Principia*, he is, most of time, actually referring to what we now call “impulse,” and he relates the imposed “motive force” to the change in “quantity of motion” (i.e., momentum) of the body.

1.11

IMPULSE AND MOMENTUM CHANGE FOR A BALL STRIKING A WALL

Example: Let us now apply our new language of impulse and momentum change to a familiar situation: A ball of mass m approaches a wall with velocity \vec{v}_1 at normal incidence (i.e., perpendicular to the wall) as shown in Fig. 1.11.1. It rebounds with velocity \vec{v}_2 after a partly inelastic collision. We take the positive direction toward the right, and we take the ball to be the system under consideration, i.e., we focus our attention on the ball alone

rather than on the ball-wall combination. The wall exerts a net force on the ball during the interaction and therefore delivers a net impulse, changing the momentum of the ball.

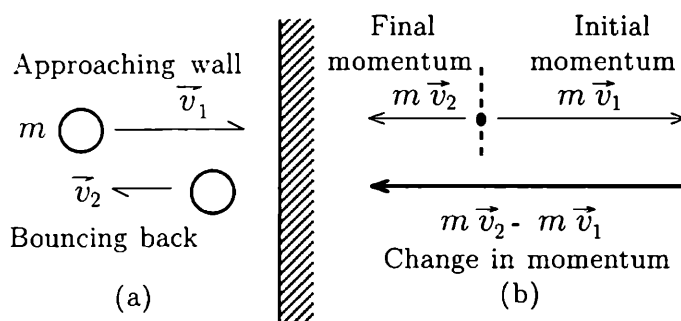


Figure 1.11.1 (a) Ball of mass m bounces from wall at normal incidence. (b) Vector diagram showing the individual momenta and the change in momentum. Positive direction is taken toward the right.

Suppose the ball has a mass $m = 100 \text{ g}$ (or 0.10 kg) and an initial horizontal velocity $\vec{v}_1 = +3.0 \text{ m/s}$. It rebounds with a smaller horizontal velocity $\vec{v}_2 = -2.0 \text{ m/s}$. We estimate the interval of contact with the wall to be approximately 0.040 s .

Let us investigate the changes that take place: The change in momentum is shown graphically in the vector diagram in Fig. 1.11.1 (b). [Be sure to draw your own diagram showing how the heavy arrow (change in momentum) is arrived at, starting with the initial and final momentum vectors.]

The numerical value of the momentum change is calculated as: $\Delta(m\vec{v}) = m\vec{v}_2 - m\vec{v}_1 = (0.10)(-2.0) - (0.10)(+3.0) = -0.50 \text{ kg m/s}$.

Interpreting this result in words: The momentum change of the ball is directed toward the left and has a magnitude of 0.50 kg m/s .

In accordance with the impulse-momentum theorem, we can now deduce the net impulse delivered by the wall to the ball. Since the net impulse has been shown to equal the change in momentum, the net impulse delivered to the ball must have been $-0.50 \text{ N}\cdot\text{s}$.

By making use of Eq. 1.9.5, we can estimate the average force exerted by the wall on the ball during the interaction:

$$\vec{F}_{\text{BA}} \equiv \frac{\Delta(m_{\text{B}}\vec{v}_{\text{r B}})}{\Delta t} \quad (1.11.1)$$

The average force must have been approximately $-(0.50)/0.040 = -12 \text{ N}$ (i.e., toward the left.) The average force exerted by the ball on the wall would have been $+12 \text{ N}$ (toward the right.)

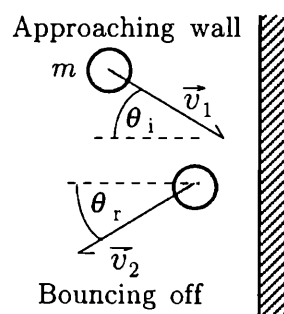
Since the final velocity is less in magnitude than the initial velocity, the collision is not perfectly elastic. The coefficient of restitution is $2.0/3.0 = 0.67$.

What would have been the momentum change and the average force on the ball if the collision had been perfectly elastic? [Ans. -0.60 kg m/s , -15 N]

Question 1.11.1 Continue the analysis illustrated in the Example above for the case in which the ball makes a perfectly *inelastic* collision with the wall, i.e., the ball is made of putty, and it sticks to the wall on striking it; in other words, $\vec{v}_2 = 0$. Take the numerical values to be otherwise the same as in the Example, but with a value of 0.20 s for the interval of interaction. (How is this interval to be interpreted now that the ball sticks to the wall? Why not say that the interval is infinite?) Be sure to re-draw the vector diagram for the momentum change. [Ans. $\Delta(m\vec{v}) = -0.30 \text{ kg m/s}$; average force on ball $\vec{F} = -1.5 \text{ N}$] How do you explain the fact that the impulse delivered to the ball by the wall is greater in any elastic (or partly elastic) collision than it is in the perfectly inelastic collision?

Question 1.11.2 A ball of mass $m = 100 \text{ g}$, moving horizontally, arrives at a wall at an angle of incidence $\theta_i = 38^\circ$ and a velocity magnitude of 3.0 m/s as in Fig 1.11.2. It undergoes a perfectly *elastic* collision and rebounds at an angle of reflection $\theta_r = 38^\circ$. The interval of contact with the wall is approximately 0.060 s .

Figure 1.11.2 Elastic collision between ball and wall in Question 1.11.2. The view is from *above*, the motion being in the horizontal plane with gravitational influence assumed to be negligible.



Analyze this collision in the same manner as we did in the example above but with due attention to the fact that we no longer have normal incidence.

- Draw the vector diagram for the change in momentum of the ball.
- What is the numerical value of the impulse delivered to the ball in the direction parallel to the wall? Explain your answer carefully.
- What is the vector momentum change of the ball?
- What is the vector impulse delivered to the ball by the wall?
- What is the average force exerted by the wall on the ball? [Answers: (b) 0; (c) -0.47 kg m/s ; (d) $-0.47 \text{ N} \cdot \text{s}$; (e) -8 N]
- Compare the vector diagrams and numerical values obtained in this example with those obtained in the case of normal incidence worked out earlier in this section. What has happened to the magnitudes of the momentum change and the net impulse? How do you explain the changes?

1.12 IMPULSE AND MOMENTUM IN A CLOSED SYSTEM

Now let us explore the connection between the impulse-momentum theorem for a single body and for the collision of two bodies comprising a closed system as in Fig. 1.12.1. Will this provide support for the momentum conservation relation that we conjectured inductively in Sect. 1.7?

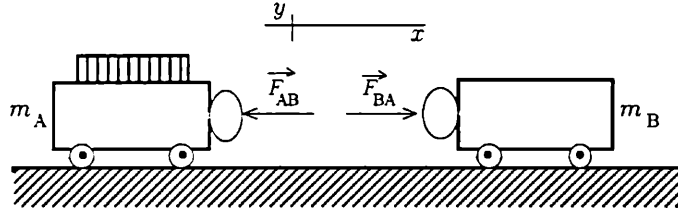


Figure 1.12.1 Force diagram for two carts undergoing rectilinear collision. Vertical forces are not shown in order to avoid clutter. Friction is considered negligible.

We apply the impulse-momentum theorem (Eq. 1.10.1) to the motion in the x direction of each of the two bodies in Fig. 1.12.1, obtaining:

$$\int_{t_1}^{t_2} \vec{F}_{AB}(t) dt = m_A \vec{v}_{A2} - m_A \vec{v}_{A1} = \Delta(m_A \vec{v}_A) \quad (1.12.1)$$

for body A and

$$\int_{t_1}^{t_2} \vec{F}_{BA}(t) dt = m_B \vec{v}_{B2} - m_B \vec{v}_{B1} = \Delta(m_B \vec{v}_B) \quad (1.12.2)$$

for body B.

Since, by Newton's third law, $\vec{F}_{AB} = -\vec{F}_{BA}$ *instant by instant* throughout the interval of interaction $(t_2 - t_1)$, the integrals on the left-hand side of Eqs. 1.12.1 and 1.12.2 are equal in magnitude and opposite in sign. If we add the two equations, the left-hand side of the resulting equation becomes zero! Let us perform the addition of Eqs. 1.12.1 and 1.12.2, obtaining:

$$0 = (m_A \vec{v}_{A2} - m_A \vec{v}_{A1}) + (m_B \vec{v}_{B2} - m_B \vec{v}_{B1}) \quad (1.12.3)$$

or, the identical idea in the alternative form

$$\Delta(m_A \vec{v}_A) + \Delta(m_B \vec{v}_B) = 0 \quad (1.12.4)$$

which says that the algebraic (or vector) sum of the momentum changes of the interacting bodies must be zero if Newton's third law holds.

Eqs. 1.12.3 and 1.12.4 can be rewritten in two alternative forms as follows:

$$m_A \vec{v}_{A2} - m_A \vec{v}_{A1} = -(m_B \vec{v}_{B2} - m_B \vec{v}_{B1}) \quad (1.12.5)$$

$$m_A \vec{v}_{A2} + m_B \vec{v}_{B2} = m_A \vec{v}_{A1} + m_B \vec{v}_{B1} \quad (1.12.6)$$

Translating these results into words: Eq. 1.12.5 says that the change in momentum of body A is equal and opposite to the change in momentum of body B. Eq. 1.12.6 says that the total final momentum of the system is equal to the total initial momentum.

The last four equations (1.12.3 through 1.12.6) are simply slightly different ways of saying exactly the same thing: namely, that, if we accept Newton's laws of motion and apply them to a collision such as that in Fig. 1.12.1, we arrive at the prediction that the final total momentum of the closed system is equal to the initial total momentum (i.e., that momentum is conserved) in collisions such as those of Fig 1.12.1 *regardless of the masses of the bodies and regardless of how the force of interaction varies during the collision.*

In Eqs. 1.12.3 through 1.12.6, we have an algebraic equation connecting the final and initial conditions without any reference whatsoever to intermediate stages or processes! The inductive guess of Huygens and his contemporaries in which we joined in Sect. 1.7 and which was based on special cases and limited empirical evidence, turns out to be fully consistent with our knowledge of dynamics as expressed in Newton's laws of motion *providing* Newton's third law holds (i.e., that the forces between objects are equal and opposite) *instant by instant* throughout the entire interval of the interaction. It is necessary to remember this very significant, "instant by instant," restriction since we shall eventually be encountering cases in which Newton's third law does *not* hold instant by instant in the interaction between separated bodies.

As we pointed out earlier, if we assume knowledge of the initial conditions (namely the masses of the interacting bodies and their initial velocities), we are still left with two unknowns (the two final velocities), and we need an additional equation to "solve the problem," i.e., to be able to predict both final conditions, given the initial conditions. As we pointed out in Sect. 1.7, we do have such an additional equation for the special case of the perfectly inelastic collision in which the two bodies stick together. It is:

$$\vec{v}_{B2} = \vec{v}_{A2} = \vec{v}_2 \quad (1.12.7)$$

Question 1.12.1 Let us examine the perfectly inelastic rectilinear collision carefully with pencil and paper:

- (a) Show that, for the perfectly inelastic collision, our preceding equations predict that:

$$\vec{v}_2 = \frac{m_A \vec{v}_{A1} + m_B \vec{v}_{B1}}{m_A + m_B} \quad (1.12.8)$$

- (b) Interpret Eq. 1.12.8 in words: (1) What happens if the bodies have equal masses and are both initially moving toward the right with A overtaking B? (2) If both bodies have equal masses but B is initially moving toward the left? (3) What happens if m_A is very small relative to m_B ? (4) If m_B

is very small relative to m_A ? (5) Make up and answer some additional questions of this variety on your own.

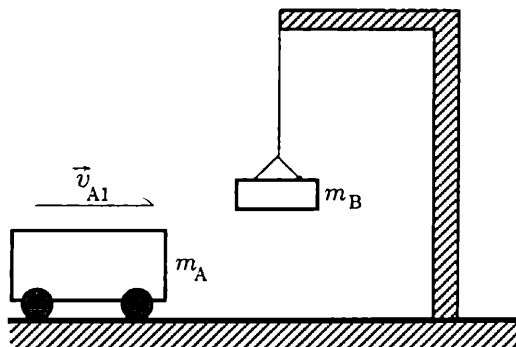
Question 1.12.2 The velocity of a bullet is sometimes determined by firing the bullet into a massive block suspended on wires and determining the velocity of the block (now containing the bullet) immediately after impact.

- Interpret this experiment in light of the discussion you have given in Question 1.12.1.
- Rewrite Eq. 1.12.8 to give an algebraic expression for the velocity of the bullet in terms of known quantities.
- Design an experiment of this variety: How might you determine the velocity of the block after impact? What might you assume for the order of magnitude of the bullet velocity and its mass? How massive would you want the block to be? etc.

Question 1.12.3 Still another form of perfectly inelastic collision is shown in Fig. 1.12.2 in which block B is dropped onto moving cart A by cutting the string suspending block B just as A passes underneath.

- What would be the algebraic expression for the velocity of the combination of B and A after the collision?
- Take the numerical case in which the mass of cart A is 3.75 kg, the mass of B is 1.25 kg, and the initial velocity of A is 68 cm/s. Calculate the velocity of the combination after B lands on the cart. [Ans. 51 cm/s]
- What role does friction necessarily play in this situation? What would happen if the interface between the block and the cart were frictionless?

Figure 1.12.2 Diagram for Question 1.12.3: The string holding block B is cut just as cart A passes underneath.



Now let us do some thinking about the perfectly elastic collision. We saw earlier that, in a perfectly elastic rectilinear collision with coefficient of restitution unity, the colliding bodies exchange relative velocities, and we therefore have an additional relation from Eq. 1.3.2:

$$\vec{v}_{A2} - \vec{v}_{B2} = -(\vec{v}_{A1} - \vec{v}_{B1}) \quad (1.12.9)$$

Question 1.12.4 Show that Eqs. 1.12.4 and 1.12.9 can be solved simultaneously to obtain the following equations for \vec{v}_{A2} and \vec{v}_{B2} in terms of the initially known masses and velocities:

$$\vec{v}_{A2} = \frac{m_A - m_B}{m_A + m_B} \vec{v}_{A1} + \frac{2m_B}{m_A + m_B} \vec{v}_{B1} \quad (1.12.10)$$

$$\vec{v}_{B2} = \frac{2m_A}{m_A + m_B} \vec{v}_{A1} - \frac{m_A - m_B}{m_A + m_B} \vec{v}_{B1} \quad (1.12.11)$$

Interpret in words what Eqs. 1.12.10 and 1.12.11 say about the case in which body B is initially at rest: Under what circumstances will body A continue moving to the right? Under what circumstances will it bounce back? How do these predictions compare with what you previously observed about rectilinear collisions?

Question 1.12.5 If the collision is *partly* elastic, we have no additional a priori relation between \vec{v}_{A2} and \vec{v}_{B2} as in Eqs. 1.12.7 and 1.12.9. What would you do in such a case, i.e., what additional information would you have to acquire in order to be able to predict one or the other of the final velocities?

1.13 THE LAW OF CONSERVATION OF MOMENTUM

We found conservation of momentum in the special case of rectilinear collisions by two different routes: (1) Induction from empirical study of simple collision experiments and application of a principle of relativity; (2) application of Newton's laws of motion to the separate bodies in simple collisions in a closed system. As we accumulate experience with such interactions in new cases, far more general and complex than the ones considered so far, we keep finding that predictions based on momentum conservation invariably turn out to be correct, and no contradictory cases have ever been encountered. In science, a regularity of such broad applicability, confirmed by extensive testing, comes to be called a law, and we shall therefore now refer to the law of conservation of momentum (abbreviated LCM for convenience in further discussion.)

What then is the status of the LCM among the regularities we observe in nature? In Sect. 1.12 we seem to have derived the LCM from the Newtonian laws of motion. Does this derivation imply that the Newtonian laws are actually more fundamental than the LCM, i.e., is the LCM simply a necessary consequence of these laws? If that is the case, we have no need of the LCM except as a matter of convenience in solving problems in which we happen to be able to connect the initial and final conditions without being concerned about impenetrable details of the history in between.

It turns out, however, that there are cases in which direct application of Newton's third law either turns out to be very difficult or, in some phenomena, impossible. In such situations, the story developed in Sect. 1.12 does not apply, and we must ask whether the LCM stands on its own and still applies

in such situations. First let us examine some of the situations in which such difficulties arise:

(1) In the Atwood machine you dealt with in your early study of dynamics, two bodies (Fig. 1.13.1), unequal in mass, are connected by a string of negligible mass which hangs over a pulley. If we hold one of the bodies, we can keep the system from accelerating. If we let go, the system accelerates. We must now examine the physics of this situation a bit more closely than we did earlier. Consider the following question: Is the time *interval* zero between the instant of our letting go the body we are holding and the instant the two bodies have the same acceleration, or does a finite time interval elapse between our letting go of the one body and the instant at which we can take the acceleration of the two bodies to be the same?

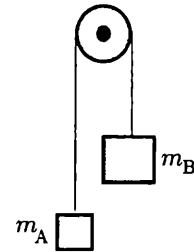


Figure 1.13.1 Atwood's machine: Two bodies connected by a very light string passing over a very light pulley (i.e., the inertial effects of the string and pulley are idealized as negligible.)

The answer here is that, during the very short but definitely non-zero time interval in which a disturbance of stretching and contracting (called a “wave”) propagates back and forth along the string, the two bodies *do not* exert equal and opposite forces on each other *instant by instant* through the string, and their accelerations are *not* equal. The condition we glibly assumed to apply in previously solving the problem does not in fact hold during the so-called “transient” disturbance that first propagates through the string and before the system “settles down” to the new, steady tension in the string. During this complicated interval, we cannot simply apply Newton’s third law to the entire system (from one body to the other) and can only apply it layer by adjacent layer in the “medium” (i.e., the string) lying between the two accelerating bodies. Overall, however, momentum is being conserved; it is, in fact, being propagated from one body to the other by the disturbance in the string.

Question 1.13.1 Consider the situations represented in Figs. 1.13.2(a) and (b): In (a) body A undergoes a collision with the left-hand end of the long coil spring which is fastened to body B at its right-hand end. In (b), body A, fastened to the left-hand end of the string, is “wiggled” up and down in a direction perpendicular (transverse) to the length of the spring. Discuss, in terms similar to those in the preceding discussion of the Atwood machine, the applicability or inapplicability of Newton’s third law directly to the interaction between bodies A and B. Is the force exerted on B equal and opposite, instant by instant, to the force experienced by A? What happens in the spring in each case?

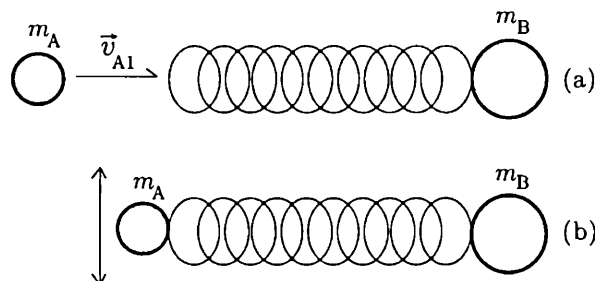


Figure 1.13.2 Interaction between two bodies separated by a spring (Question 1.13.1).

Question 1.13.2 A diaphragm clicks at one end of a room, making a sound as it accelerates. At the other end of the room, the sound is detected by another diaphragm, that of a microphone. Do the two diaphragms exert equal and opposite forces on each other instant by instant in accordance with Newton's third law? What is it that happens physically in this interaction?

(2) Two charged particles separated by a distance r_1 in a vacuum, as in Fig. 1.13.3, are initially stationary and repel each other with equal and opposite forces in accordance with Newton's third law.

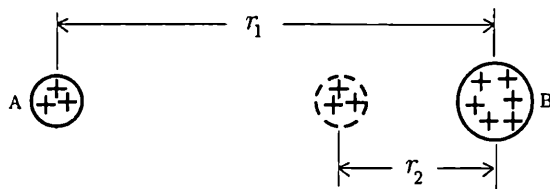


Figure 1.13.3 Positively charged particles A and B in a vacuum. Particle A is suddenly displaced from position r_1 to position r_2 relative to particle B.

Particle A is suddenly displaced toward particle B from position r_1 to position r_2 . At the latter position, particle A must experience a larger repelling force than it did at position r_1 , and the force must have been changing continually as A was being displaced. What must have been happening to particle B? Did the force at B change at the instant the displacement of A began? Does Newton's third law apply to this situation instant by instant?

This is not a question that has an immediate and simple pat answer. It is a question that concerned 19th century investigators of electric and magnetic phenomena, especially Faraday and Maxwell, and it took a long time to resolve. The resolution resided in Maxwell's invention of the theory of the electromagnetic field. Maxwell showed that electromagnetic changes or disturbances are propagated through a vacuum in a manner analogous to that in

which disturbances are propagated through springs and sound waves are propagated through air. In other words, Maxwell showed that Newton's third law does *not* apply instant by instant in the phenomenon of Fig. 1.13.3, but that the law of conservation of momentum is still obeyed. In this instance there is no material medium between the two particles, yet a disturbance (obeying conservation of momentum) is known to be propagated from one point to another at a finite, albeit extremely high, velocity, namely the velocity of light. (You will have opportunity to learn more about such matters when you study electricity and magnetism.)

A question similar to that of Fig. 1.13.3 applies to the interaction between gravitating bodies. Newton assumed that the third law was fully valid in this case, and the Newtonian theory of gravitation is thus called an "action at a distance" theory, the term "action at a distance" referring to the instant by instant requirement of equal and opposite forces between interacting bodies. Although this view is fully adequate for most astronomical calculations, it is now believed (in the light of Einstein's general theory of relativity) that the physics of gravitation *does* involve wave propagation analogous to that in electromagnetic phenomena. Thus the general theory of relativity and Maxwell's electromagnetic theory are *not* action at a distance theories. They abandon this part of the Newtonian scheme.

The various cases we have been examining in this section illustrate the fact that, in physical theory, we are eventually forced to abandon the Newton's third law requirement that separated bodies always exert equal and opposite forces on each other instant by instant. It turns out, however, that we can always save the day by turning to what emerges as a deeper and more general regularity, namely that momentum is conserved—and is transported from one body to another through intervening space. The law of conservation of momentum is therefore seen as being more deeply fundamental than the Newtonian theory even in "classical" physics. In "relativistic" physics (at very high velocities, approaching that of light), Newtonian mechanics fails completely, and the theory of relativity turns to the law of conservation of momentum as one of its fundamental underpinnings.

All this does not mean, however, that Newtonian mechanics is a deficient theory that must be thrown away. $\vec{F}_{\text{net}} = m\vec{a}$ and the third law are powerful and wonderfully useful tools within their very broad range of applicability to many terrestrial and celestial phenomena. As with most tools, literal and figurative alike, it is necessary to know when and under what circumstances it is *not* appropriate to use them.

The intent of this section has been to enlarge your perspective toward the physical concepts we have been developing and to help you to do qualitative physical thinking that will prepare you for better understanding of the more advanced theories you will encounter later in your study of physics.

1.14 CENTER OF MASS OF A SYSTEM OF PARTICLES

Up to this point, we have been thinking about the law of inertia (Newton's first law of motion) in terms of the behavior of a single particle. Let us state this law in language that will be helpful in the present context: A particle moves in a straight line at uniform velocity unless an interaction with another object alters either the magnitude or the direction (or both) of the velocity. When we discern a change in the velocity of our particle, we infer the presence of such an interaction, and we say that an external force is acting on the particle.

Now that we have investigated the behavior of systems of interacting particles and developed a foundation for accepting the general law of conservation of momentum for a closed system, it becomes appropriate to ask whether or not some regularity in velocity, similar to that described by the law of inertia, is still to be discerned in the behavior of an entire system. It turns out that we can always identify a point (in any closed system of particles) that continues moving in a straight line at uniform velocity unless an external force acts on, and accelerates, the system. This point, incidentally, may be located in empty space among the interacting bodies and is not necessarily anchored in a material object. In the following discussion we shall identify the point in question.

We start, as we have done so frequently in the past, by working out a simple special case, and we shall then proceed to generalize from this case. The special case is that represented in Fig. 1.14.1 in which two particles with different masses are moving at different velocities along the x -axis of our frame of reference.

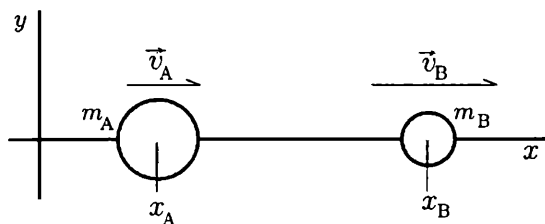


Figure 1.14.1 Particles of mass m_A and m_B , at instantaneous positions x_A and x_B , moving along x -axis at velocities \vec{v}_A and \vec{v}_B respectively.

Let us represent the total momentum of any system by the symbol \vec{P} , this total being the vector sum of the individual momenta of the constituents. The law of conservation of momentum says that the *total* momentum \vec{P} of a system remains constant (i.e., unchanging in either magnitude or direction), regardless of the motions and interactions that take place inside, as long as no material goes in or out and as long as there is no net, unbalanced force acting on the system from the outside.

Let us denote the total mass (i.e., the sum of the individual masses) within our system by $\sum m$ (i.e., $\sum m \equiv m_A + m_B$.) We shall now try to identify a

single value of velocity that characterizes the whole system in such a way that the actual total momentum \vec{P} is equal to the simple product of the total mass of the system and this special, representative velocity (analogous to the product for the momentum of a single body.) If we denote this representative velocity by \vec{v}_{cm} , we say that we are looking for the value of \vec{v}_{cm} that makes

$$\vec{P} = (\sum m)\vec{v}_{\text{cm}} \quad (1.14.1)$$

(Why we choose the subscript cm will emerge as we go along.)

In other words, \vec{v}_{cm} is some sort of average velocity, one that gives the total momentum of the whole system when it is multiplied by the total mass. We shall identify the nature of the average when we find out more about this representative velocity.

For the special case of Fig. 1.14.1, with particle velocities confined to the x -axis, we shall drop the vector arrow notation in order to reduce clutter, and we shall denote positions along the x -axis by the symbol x with their implied algebraic signs and velocities of objects along the x -axis by v_x , also with their implied algebraic signs. We then have, for Fig. 1.14.1, the relations:

$$\sum m = m_A + m_B \quad (1.14.2)$$

and

$$P_x = (\sum m)v_{x\text{ cm}} = m_A v_{x\text{ A}} + m_B v_{x\text{ B}} \quad (1.14.3)$$

where the v_x 's represent *algebraic* values of velocity along the x -axis.

Combining Eqs. 1.14.2 and 1.14.3 gives

$$v_{x\text{ cm}} = \frac{m_A v_{x\text{ A}} + m_B v_{x\text{ B}}}{\sum m} = \frac{m_A v_{x\text{ A}} + m_B v_{x\text{ B}}}{m_A + m_B} \quad (1.14.4)$$

and we should keep in mind the fact that

$$v_{x\text{ A}} \equiv \frac{dx_A}{dt} \quad \text{and} \quad v_{x\text{ B}} \equiv \frac{dx_B}{dt} \quad (1.14.5)$$

The quantity $v_{x\text{ cm}}$, is, as we said earlier, an *average*. It is called a “weighted average” since it is “weighted” in Eq. 1.14.4 by the values of the masses of the individual bodies.

Question 1.14.1 Examine the algebra of Eq. 1.14.4 and show:

- (a) That $v_{x\text{ cm}}$ is half way between the two individual velocities (i.e., half the sum of the two velocities) if the masses are equal.
- (b) That $v_{x\text{ cm}}$ is nearer to the velocity of body A if m_A is larger than m_B , and
- (c) That $v_{x\text{ cm}}$ is nearer to the velocity of body B if m_B is larger than m_A .

This analysis illustrates the *meaning* of the term “weighted average.”

The average velocity $v_{x\text{ cm}}$ should be the velocity of some point within our system whether or not there is a material object at this point. Let us call the location of this point x_{cm} . The connection between $v_{x\text{ cm}}$ and x_{cm} is, of course, given by the derivative

$$v_{x\text{ cm}} = \frac{dx_{\text{cm}}}{dt} \quad (1.14.6)$$

Combining Eqs. 1.14.4, 1.14.5, and 1.14.6, we obtain

$$v_{x\text{ cm}} = \frac{dx_{\text{cm}}}{dt} = \frac{m_A \frac{dx_A}{dt} + m_B \frac{dx_B}{dt}}{m_A + m_B} = \frac{d}{dt} \left[\frac{m_A x_A + m_B x_B}{m_A + m_B} \right] \quad (1.14.7)$$

Comparing the second and last terms in Eq. 1.14.7, we note that the quantity in the brackets must be equal to x_{cm} , and, like $v_{x\text{ cm}}$, it represents a weighted average—in this case a weighted average of the *positions* of the two bodies along the x -axis. As a consequence, we can now say that

$$x_{\text{cm}} \equiv \frac{m_A x_A + m_B x_B}{m_A + m_B} \quad (1.14.8)$$

In other words, x_{cm} is that position along the x -axis which moves with the average velocity $v_{x\text{ cm}}$ whether or not there is any object at this location.

Now let us go back to the LCM, which says that the total momentum must be constant for any closed system. This means that momentum changes of individual objects within the closed system (owing to internal collisions or to internal attractive or repulsive forces) must add up to zero. Thus, with P_x and $\sum m$ both constant, $v_{x\text{ cm}}$ must also be constant (see Eq. 1.14.3) for the closed system. If $v_{x\text{ cm}}$ is constant, the point x_{cm} must move at uniform velocity. In other words, the particular point x_{cm} within the closed system moves in a straight line at uniform velocity regardless of how objects within the closed system might be changing their motions. The point x_{cm} is given the name “center of mass of the system,” and we shall, from time to time, use the abbreviation cm for this point, as we have already anticipated in our subscript notation.

Our argument has focussed on the special case in which two objects move and interact along the x -axis, but this argument is easily extended to two objects moving in the x , y , and z directions and then to any number of bodies moving in three dimensional space. A dramatic illustration of the implication of this generalization is that, if a single body moving relative to us in a straight line at uniform velocity breaks up (due, say, to an internal explosion), and its parts spread out in all directions, the *center of mass* of the collection of parts keeps on moving relative to us at the same velocity as that of the original single object!

Question 1.14.2 Given Eq. 1.14.8, what would be the expression for the location of the center of mass if one of the two objects were placed at the origin of the coordinate system? Interpret the expression by finding where the cm would be located if the masses were equal and where it would be located if either one of the masses were much larger than the other.

Question 1.14.3 Suppose the distance between the two bodies is denoted by R . Show that the distances r_A and r_B of each of the bodies from their center of mass, respectively, is given by

$$r_A = \frac{m_B}{m_A + m_B} R \quad \text{and} \quad r_B = \frac{m_A}{m_A + m_B} R \quad (1.14.9)$$

and that the ratio of the two distances from the cm is equal to the *inverse* ratio of the masses, i.e.

$$\frac{r_A}{r_B} = \frac{m_B}{m_A} \quad (1.14.10)$$

Interpret these results by describing what happens to r_A , r_B , and their ratio as the masses of the two objects change relative to each other from large, through equal, to small values.

Question 1.14.4 Choose some numbers of your own for masses and positions in Fig. 1.14.1, and calculate the location of the center of mass (cm) of the system. Play with the numbers so as to see how the location of the cm varies as one mass is increased or decreased relative to the other.

Question 1.14.5 The moon has a mass 0.0123 that of the earth, and the distance between the centers of the earth and the moon is equal to 60.3 earth radii (R_E). Find the location of the center of mass of the earth-moon system. [Ans. $0.729R_E$ out from the center of the earth (i.e., within the earth itself.)]

Question 1.14.6 What would be the form of Eq. 1.14.8 if there were three bodies instead of two moving along the x -axis as in Fig. 1.14.1? What would be the form if there were n bodies so moving?

Question 1.14.7 Suppose that the two bodies in Fig. 1.14.1 had coordinates x_A , y_A and x_B , y_B respectively in the x - y plane instead of being confined to the x -axis. What would be the expressions for the coordinates (x_{cm} , y_{cm}) of the center of mass of the system?

Question 1.14.8 Sketch how each of the collisions we dealt with in Sects. 1.4 and 1.5 would look from a frame of reference moving with the center of mass of the system of the two bodies.

1.15 QUESTIONS AND PROBLEMS

1.15.1 When, in Sect 1.11, we dealt with the momentum changes of balls bouncing from a wall, we were analyzing the impulse and momentum changes of a single object (the ball) and did not examine the effects from the standpoint of a closed system. The relevant closed system in these cases is the combination of the ball and the entire

earth to which the wall is fastened. Let us now think through what happens in this closed system.

- (a) Using the numerical results obtained in the various examples dealt with in Sect 1.11, discuss what happens to the entire earth as the ball undergoes its momentum changes as a result of impulse delivered by the wall. How do you account for the fact that we cannot possibly detect the effect on the earth?
- (b) Without going into numerical calculations, describe in detail the momentum changes within the system consisting of the ball and the earth when you throw the ball. (Take the ball as one of the interacting objects and the combination of you and the earth as the other.) How do you account for the fact that the velocity imparted to the earth is negligible?
- (c) Describe in detail the sequence of momentum changes and impulses delivered to both you and the earth when you go through the act of jumping vertically upward and coming back down for a perfectly inelastic collision with the earth. (Keep in mind the fact that, because of the gravitational interaction, you are interacting with the earth even when you are on the way up and not in contact with it.)

1.15.2 Imagine yourself supplied with a cart that rolls with negligible frictional effects and is large enough for you to walk about on. (A boat on the water would do equally well.) Denote your own mass by m_S and that of the cart by m_C , and adopt appropriate symbols for velocities relative to the laboratory frame of reference. Make *algebraic* analyses of the following situations:

- (a) Suppose that you are initially standing at one end of the cart and that all velocities relative to the chosen frame of reference are zero. You walk to the other end of the cart and stop. Describe in detail what happens: What is your final displacement relative to the cart? Relative to the ground? What is the displacement of the center of mass of the system? What is the sequence of momentum changes of each object? What is the total momentum of the *system* during this experiment? What are the final velocities of you and the cart? (In your analysis, consider various possible ratios of the masses, i.e., what happens when $m_S/m_C < 1$?, when $m_S/m_C = 1$?, when $m_S/m_C > 1$?)
- (b) Suppose you run toward the initially stationary cart and jump on to it with a horizontal velocity v_{S1} relative to the ground. You rest for a moment and then walk to the far end of the cart and jump off with a horizontal velocity (relative to the ground) greater than v_{S1} . Analyze and describe the sequence of momentum changes. What are the final velocities of the objects in the system? Has momentum been added to the system by your jumping off with a higher velocity than that with which you jumped on?
- (c) Invent some other possible experiments and predict the pertinent results.

1.15.3 Although air resistance is not really negligible in the following situation, let us take it as negligible in the first approximation. Consider the case in which a shell, projected from a cannon, is moving in its normal parabolic path characteristic of projectile motion. The shell explodes in midair, and the fragments spread out in all directions. The system of fragments is not a closed system since all the fragments are interacting gravitationally with the earth just as the whole shell was doing before the

explosion. What will be the trajectory of the center of mass of the fragments after the explosion? Explain your reasoning carefully in your own words.

1.15.4 Describe in words, in terms of momentum changes of various objects, the mechanism of propulsion of a propeller-driven boat or airplane and of a rocket. Point out carefully the similarities between the two cases as well as the essential difference.

1.15.5 In a laboratory experiment, a marble (mass = 15.0 g) struck horizontally near the edge of a table flies off the edge of the table with a horizontal velocity v_{x1} . The edge of the table is 80.0 cm above the floor, and the marble lands at a horizontal distance of 1.85 m from the edge of the table. (Sketch a diagram of the situation just described.)

- Calculate the initial momentum of the marble and the momentum at the instant it strikes the floor. (Remember that momentum is a vector quantity. You will find it best to keep track of the x and y components separately.)
- Draw a vector diagram showing the change of momentum of the marble between the instant it flies off the table and the instant it arrives at the floor, and calculate the numerical magnitude of this change.
- What is the magnitude and direction of the impulse delivered to the marble during its flight? (Evaluate the impulse in two different ways and make sure they check.)

Partial answers: $|mv_{x1}| = 0.0687 \text{ kg m/s}$; $|\Delta(m\vec{v})| = 0.0594 \text{ kg m/s}$.

1.15.6 A stream of particles, each with mass m and velocity v , strikes a wall at normal incidence (i.e., the particles arrive from a direction perpendicular to the wall.) Suppose that the stream is such that N particles arrive at the wall in each second.

- First, argue from the relation between impulse and momentum that the average force exerted on an object must be equal to the rate at which momentum of the object was changing, i.e., the amount by which the momentum changes in each succeeding second. Then obtain algebraic expressions for the average force exerted on the *wall* by the continuing stream of particles if their collision with the wall is (1) perfectly *inelastic* and (2) perfectly *elastic*. Be sure to maintain a clear distinction between the impulse being delivered to the collection of particles and the impulse being delivered to the wall. [Ans. (1) Nmv , (2) $2Nmv$]
- The stream of particles strikes the wall over an area A . What is the algebraic expression for the average *pressure* exerted on this area? (“Pressure” is the name for “force per unit area.”) If the mass of each particle is 2.0 g, the velocity is 8.0 m/s, 50 particles are arriving at the wall in each second, and the area on which they strike is 12 cm^2 , what is the pressure on the wall if the particles are making perfectly elastic collisions? How does this pressure compare in magnitude with atmospheric pressure? (Look up the latter value in order to make the comparison.) [Ans. 1300 N/m^2]

1.15.7 In his *Astronomia Nova*, of 1609, seventy eight years before Newton's *Principia*, Johannes Kepler made what has become a famous and often-quoted remark: "If two stones were placed in any part of the world, near each other yet beyond the sphere of influence of a third related body, the two stones, like two magnetic bodies, would come together at some intermediate place, each approaching the other through a distance in proportion to the mass of the other." In other words, Kepler was saying that ratio of the two displacements would vary inversely as the ratio of the masses. (Kepler's word for mass was then the Latin "moles"—a term denoting bulk of matter in some vague sense rather than our modern, operationally defined concept.)

Clearly defined conceptions of inertial and gravitational mass were then still far in the future; yet Kepler must be given credit for a profound insight. Comment on the statement quoted above: In modern terms, what are the dynamical implications? Is the prediction consistent with our present knowledge? Why or why not? What connection do you see between Kepler's statement and the results you obtained in Question 1.14.2? What name have we given the "intermediate place" to which Kepler refers? What relation do you see between this situation and that of the so-called "reaction car" experiment in the laboratory? (The "reaction car" experiment is that in which two carts on a table, or gliders on an air track, have a compressed spring between them and fly apart when the spring is released.)

1.15.8 Two carts (or two gliders on an air track), carrying springs as in Fig. 1.2.1 (a), are held together with the springs compressed and are allowed to spring apart after being released from rest. Take the mass of cart A to be 1.50 kg and that of cart B to be 0.50 kg. The final velocity of B is observed to be +2.3 m/s. Indicating what you take to be your system and explaining your reasoning, calculate the final velocity of cart A and the final velocity of the center of mass of the system.

1.15.9 A time varying force, described by the function $F_B(t) = 45(2t - t^2)$, acts on body B during the interval $0 \leq t \leq 2.0$, where F_B is expressed in newtons and t in seconds.

- Sketch the F_B versus t history of variation of the force.
- Calculate the impulse I delivered to the body by F_B and calculate the average value \bar{F}_B of F_B .
- Taking F_B to be the *net* force acting on B and assuming the mass of the body to be 2.8 kg, calculate the velocity change imparted to B.

Answers: $I = 60 \text{ N} \cdot \text{s}$; $\bar{F}_B = 30 \text{ N}$; $\Delta v_B = 21 \text{ m/s}$.

1.15.10 The net force acting on a glider on an air track (essentially frictionless system) varies with time as shown in the following figure. The glider has a mass of 0.850 kg. When the force is suddenly applied to the glider at clock reading $t = 0.0 \text{ s}$, the glider has an instantaneous velocity of 0.150 m/s in the positive direction.

Chapter 2

Interaction, System, State, and Conservation of Mass

2.1 INTRODUCTION

In observing changes taking place in the world around us, we see, over and over again, that change of one kind or another results from the *effect* that objects (or groups of objects) seem to have on each other. The objects (or groups of objects) show evidence of “influencing” or “doing something” to each other. We invented the concept of “force” in connection with the effect exerted by one object on another when acceleration is being imparted or when objects are being deformed. We recognized the rubbing effect to which we give the name “friction” as a force exerted by one object on another when there is sliding contact at an interface. In the preceding chapter we dealt with momentum changes that were produced when objects collided with each other. We recognize the effect of moving water and air in the erosion of soil and rocks. We observe that, when objects at two different temperatures are brought in contact with each other, the temperature of the higher temperature object decreases while that of the lower temperature object increases until the changes cease when the temperatures become equal. We observe that many materials become electrically charged when rubbed against each other.

When we observe changes taking place through the effect or influence that one object or group of objects seems to have on another object or group, we say that an “interaction” has taken place, and we shall now extend this concept beyond the very narrow range of phenomena we have discussed earlier.

We give the name “mechanical” to interactions such as those in which we accelerate a block along the floor, perceive objects changing their motion on colliding with each other, or bend a rod by applying forces to it. We give the name “thermal” to interactions in which contact between objects at different temperatures results in temperature changes of the objects or in effects such as melting or freezing. When a metal dissolves in an acid with liberation of

hydrogen gas, when wood or paper burn in air, when carbon dioxide turns lime water milky, we say that “chemical interaction” has taken place. When objects that have been rubbed with some other material attract or repel each other or attract bits of paper, we speak of “electrostatic interaction.” When magnets attract pieces of iron or attract or repel each other, we speak of “magnetic interaction.” When we deal with the forces that the earth and all other objects (including ourselves) exert on each other or with the forces between the sun and the planets in the solar system, we speak of “gravitational interaction.”

In each instance we have been illustrating, we can describe the specific changes that have taken place between the observed initial and final conditions, and we cite these changes as *evidence* of the interaction. The changes may be changes in motion, deformation, density, temperature, pressure, chemical, electrical, magnetic, or any other properties.

In many instances, of course, more than one type of interaction may be occurring simultaneously. In the case of acceleration of the bodies in an Atwood machine, the interaction between the two bodies via the connecting string is mechanical, while that between the earth and the bodies is gravitational. When the red powder mercuric oxide is decomposed into liquid metallic mercury and gaseous oxygen by heating over a flame, chemical and thermal interactions are both taking place. When water is decomposed into hydrogen and oxygen by passage of an electric current, both chemical and electrical interactions are taking place.

There are still other kinds of interaction that you will encounter in further study. Interactions in which electric and magnetic effects are inextricably intertwined are called “electromagnetic.” Light, for example, turns out to be an electromagnetic phenomenon, and its interaction with matter is also electromagnetic. So are interactions among atoms and molecules. On a deeper microscopic level we encounter “nuclear” interactions, and so forth.

Notice that, in science, we use the term “interaction” in connection with observed changes in which we feel we have clear cut *evidence* that one object (or group of objects) has “done something” to another object or group. Such evidence resides, for the most part, in our ability to reproduce the effect under similar conditions at different times and places, and to vary the magnitude of the effect by changing various properties of the interacting objects (such as their motion or temperature or concentration, etc.) In other words, evidence of interaction involves clear and controllable causes and effects.

Some people claim the existence of interactions for which such clear cut evidence does not exist [e.g. witchcraft, mind reading, telepathy, changing the motion of objects by willing the change (psychokinesis), astrological effects, bringing about rain by executing a rain dance], but such “interactions” are not accepted within the bounds of natural science even though many such effects were accepted at various times, in various cultures, over the history of evolution of human thought, and, in some quarters, are accepted even now.

2.2 SYSTEM

As we start discerning order and connections in what initially appears to be a chaotic and disordered flux of natural phenomena, it becomes convenient to focus our attention on specific objects (or groups of objects) and to give detailed descriptions of their behavior, condition, and change of condition with time. In creating the sciences of kinematics and dynamics, we focussed our attention on a single particle and on its interaction with the earth or with some other object (a string, the floor, a colliding body, our hand) that imparts a mechanical force. In analyzing the Atwood machine, we were concerned with two particles, each interacting mechanically with a string (and with each other through the string), while each particle also interacts gravitationally with the earth. In examining the effects of collision between two objects, we, on some occasions, focussed our attention on one of the objects at a time, but, on other occasions, we dealt with both objects simultaneously as a group. What object or group we select for attention or description in any given circumstance is completely a matter of choice on our part. The choice, however, is frequently dictated by convenience, or by elegance and simplicity, or by the power of the description or insight being generated.

We give the name “system” to whatever object or group of objects we select for our attention, and it is essential, both for ourselves and for anyone with whom we wish to communicate, that we always specify very clearly and explicitly just what object or group comprises the system we have chosen. (Note that we have already been using the term “system” in the preceding chapter, but we deferred standing back and examining its meaning explicitly.)

In dealing with the Atwood machine, we might take the string and each one of the two suspended objects as three separate systems, writing a separate dynamical equation for each one of the three. On the other hand, we sometimes take the entire assembly of the three objects as our system and deal directly with the gravitational interaction between this system and the earth. In dealing with a car moving along the road, we might want to take the entire assembly (engine, car, driver) as the system and deal with the frictional interactions between this system and the road and the surrounding air. In another instance, we might wish to deal with the engine as the system and consider its interaction with the rest of the car.

In some instances we might want to examine interactions that take place inside a system we have been considering. We might then identify a “subsystem” on which to focus attention (e.g., the engine within the car as a whole, or the wheels that interact with the road on the one hand and with the car on the other). We might find ourselves needing to deal with chemical interactions between parts of a system we have chosen initially or with thermal interactions between parts of such a system.

The points to keep in mind are the following: (1) We are taking the common words “system” and “interaction,” separating them from everyday speech,

and adopting them as names for new scientific concepts that we have just invented for our use and convenience in describing natural phenomena. (2) The choice of systems and subsystems in any given situation is entirely up to *us* and is not imposed automatically. We must therefore always make explicitly clear to ourselves, and to anyone else, precisely what we are taking to be our systems and subsystems in any given discussion.

Question 2.2.1 Consider the following situations. In each one identify the evidence for occurrence of interaction, indicating what specific changes you discern, and, if there is no evidence of interaction, say so explicitly. Define systems that you might take as interacting with each other and name the interactions you can identify. (Note: This question is quite open ended, and there is no one, “correct,” pat answer for each case. That makes it a good question to discuss with fellow students. You must exercise your own judgment. Answers depend on your choice of systems and on your own decisions on whether, for example, to include or exclude interaction with the earth. Keep in mind the fact that evidence for gravitational interaction between systems in static situations is not obvious from changes taking place but must be inferred from other relevant experience.)

- (a) A cart is let go at the top of a sloping plank.
- (b) A person walking along the street steps on a slippery spot and falls to the ground.
- (c) A tightly *covered* container of water stands in a room.
- (d) An *open* container of water stands in a room.
- (e) A toaster is plugged into an electric wall outlet.
- (f) A stone is thrown into a still pond.
- (g) An aluminum pot stands on a kitchen shelf.
- (h) An iron pot rusts as it stands on a kitchen shelf.
- (i) Air rises in the neighborhood of a hot radiator in a room.
- (j) Water boils in a pot on the stove.
- (k) An initially warm house cools down in cold weather when the heating system is turned off.
- (l) A planet revolves around the sun.
- (m) You experience a shock in dry weather after scuffling over a rug and bringing your finger near a doorknob.
- (n) Standing on the ground, you jump vertically upward.

2.3 STATE OF A SYSTEM AND CHANGES IN STATE

In describing changes that are evidence for interaction between systems, we must refer to various *properties* of the systems in question. For example, we note the number of objects, changes in velocity or momentum, changes in elevation relative to the surface of the earth, changes in size or shape, changes

in mass, changes in temperature, pressure,¹ density, chemical composition, electric or magnetic condition, solid or liquid or gaseous condition, and so forth. A list of relevant properties is needed to specify completely the condition of the system (one object or a group of objects) at any instant. Given such a complete list, we could reproduce the system from scratch if we wished to do so. When we have such a list, we say we know the “state” of the system. Properties that change as the state of a system changes and that are fixed when the state is fixed are called “state variables” or “functions of state.”

The state of a system can change through internal interactions among its own subsystems (e.g., collision between two objects when the two objects constitute the system; a liquid within an insulated container coming to uniform temperature after starting with higher temperature at the bottom than at the top) or it can change through external interactions with another system. Change of state is prime evidence of interaction of one kind or another. One must be careful, however, about interpreting an *unchanging* state. This *might* imply absence of interaction with another system, but a fixed state might also be maintained by interactions that just compensate each other (e.g. an object moving at uniform velocity under the influence of two balanced forces; a chunk of metal retaining a constant temperature distribution along its length through being in contact with a higher temperature body at one end and a lower temperature body at the other).

Question 2.3.1 What variables must be specified to define completely the state of an Atwood machine during some instant while it is accelerating?

Question 2.3.2 Make up some additional illustrations of

- (a) systems that exhibit changes of state through interactions among their subsystems.
- (b) systems that do not change state while still undergoing interactions with other systems.

In Chapter 1, where we were concerned only with velocity and momentum changes of colliding bodies, the state of any given system, at any instant of time, was adequately specified by the numerical values of masses and instantaneous velocities. It turns out, however, that, in inelastic collisions, the temperature of the interacting bodies is always found to rise. This implies an additional change of state that we have, so far, not taken into account. We shall soon see that it is necessary to pay attention to such changes in order to attain a more complete description of the interactions taking place than we achieved through recognizing conservation of momentum alone.

In order to form a better idea of just what we mean by specifying the state of a system, we shall consider a few examples beyond the purely mechanical

¹“Pressure” is a subtle concept that has not yet been properly defined in our sequence of development. We shall come back to it and develop its meaning more carefully in the following section.

ones encountered in Chapter 1. Let us start with situations in which we are not concerned with motion, or changes in motion, or changes in elevation, as, for example, a system consisting of some amount of a single pure material such as iron, or aluminum, or water, or alcohol, or carbon dioxide, or nitrogen. (The gases and liquids would, of course, be held in a container which we would not consider part of the system at this point.) It is an *observed fact* (not something that is figured out from more basic principles) that the state of such simple systems is completely fixed if we specify the mass of the given material and its pressure and temperature. If, without changing the total mass, we change either the temperature of the material, or the pressure under which our sample is confined, or both, we find that the state of our sample changes as evidenced by the fact that the *volume* of the sample changes. (While the volume changes of gases under such circumstances are quite obvious, the volume changes of liquids and solids are usually quite small and may be difficult to observe without using fairly delicate instruments, but *all* materials exhibit changes in volume under changes in temperature and confining pressure, except in very special circumstances in which the effects of simultaneous pressure and temperature changes happen to balance each other.)

In many instances, especially when we are not concerned with changes in motion of the system as a whole, we may not be interested in its total mass, and we focus our attention only on the condition of the material in question. Then, instead of talking about either the total mass or total volume, we simply talk about the *density* of the material, i.e., the mass in one unit of volume as in g/cm^3 or kg/m^3 . (Density is another state variable, one that we have not had occasion to mention up to this point.) When the temperature and pressure of a pure material are specified, it turns out that the density is fixed; we cannot specify the density independently. If we do, for example, specify the density and the pressure, it turns out that this combination is possible at only one particular temperature. In other words, specifying any two state variables in the case of a single pure material, fixes the state of the material completely except for the total amount of the material we might have. In specifying the state of systems, we usually specify the state variables temperature and pressure because these variables are most directly under our control. We usually take them to be our “independent” variables, not in the sense that they are beyond our control, but in the sense that other properties of the system (such as density) are *dependent* on the temperature and pressure, and the latter *are* under our control.

Variables such as total mass and total volume of a system, which depend on how much material we happen to have, are called “extensive” variables, while variables such as temperature, pressure, and density, that do not depend on how much material we have and indicate the condition of the material regardless of amount, are called “intensive” variables.

Remember that, in saying that the state of a single pure material is completely fixed when we specify the two independent variables pressure and tem-

perature, we are asserting an *empirical fact* (i.e., one based only on observation and experiment); this fact cannot be derived from some fundamental theory.

To specify the state of a more complex system, we must specify additional variables. The next level of complexity is, for example, a system consisting of two materials instead of just one. (We say that the system has two “components.”) In such a case the state turns out to be fixed if we specify three variables, namely temperature, pressure, and *composition*. Composition can be specified, for example, by giving the ratio of the masses of the two components that are present, or, what is essentially the same thing, the percentage, by mass, of each.

Question 2.3.3 Suppose we have a mixture of the gases oxygen and nitrogen in a cylinder.

- (a) What variables would you specify in order to fix the state of the system?
- (b) If you are told that the cylinder contains air at a pressure of 30 lb/in² and a temperature of 25°C, is the state of the gas fixed? Why or why not?

Question 2.3.4 Consider a water solution that contains both ordinary salt (sodium chloride) and magnesium sulfate. What would you have to specify in order to fix the state of the system?

There are many other changes in state that we shall not undertake to consider in any detail at this point, but let us note a few examples to help maintain our perspective. Materials can exist as solids, liquids, or gases (as in the case of water ice, liquid water, and water vapor). We speak of these as different “phases” of the material. Different phases exist at different combinations of pressure and temperature, and two phases may coexist under special conditions, as ice and liquid water coexist at 0°C and atmospheric pressure. We also change the state of systems by subjecting them to electric or magnetic fields.

Our principal concern with changes of state will turn out to have to do with the concept of “energy” which will be developed in Chapter 4. All changes in state stem from interactions of one kind or another (either between separate systems or among subsystems within a system), and all such changes involve what we come to call “energy transformations.” One of the great discoveries of physics during the 19th century was that it is possible to find ways of calculating numbers that are preserved in such interactions—numbers that obey a conservation law, different from and in addition to, the law of conservation of momentum.

2.4 PRESSURE

We have mentioned “pressure” several times as being a crucial variable in describing the state of a system when we are in need of more detail than just

the masses, or velocities, or temperatures of its parts. We must recognize that, so far however, we have been talking rather loosely since we have not given the term “pressure” clear technical meaning. As in many other instances, we are taking a word out of everyday speech and making it a scientific *concept* with a technical meaning that may be distantly related to its everyday meaning but with profound differences that must be understood. Technical terms require careful *operational definition*, that is, we must describe exactly what we *do* in order to assign numerical values to the concept. Pressure, the idea we are now getting at, does have to do with squeezing, compressing, and expanding, as we would suppose given the everyday meaning of the word, but now let us proceed with the extension and refinement.

Consider a solid rectangular block of material such as wood or metal with length, width, and height all having different values (as in Fig. 2.4.1). We can place the block on the table on any one of three different faces, each face having a different area. Now consider the upward force exerted by the table on the block—the force that is equal and opposite to the downward pull of the earth on the block. Up to now you have probably thought of this as a single force acting upward on the block without worrying about how this effect is produced, and there is nothing wrong with having thought in this simplified way since deeper analysis was not relevant to the problems you were considering. Now, however, we must look more deeply.

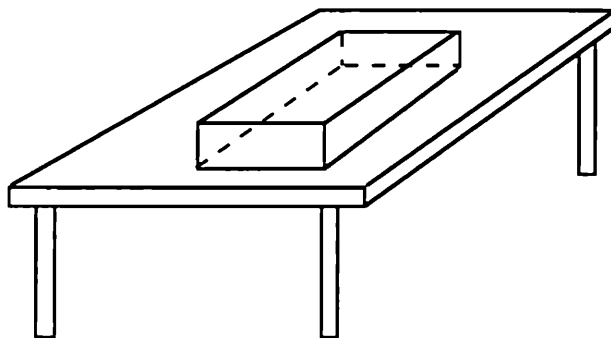


Figure 2.4.1 Solid rectangular block lying on a table.

The total downward pull of the earth on the block (i.e., the weight of the block) is not really one single force acting at a single point. Its single value conveniently represents the resultant of a “smeared out” effect, being equal to the sum of all the little downward pulls, added up chunk by chunk (i.e., integrated), over the entire block. In a similar way, the total upward force exerted by the table on the block is also the resultant of a smeared out effect. It is equal to the sum of all the little upward forces acting on each little area into which we can imagine dividing the bottom face of the block. The total upward force is the same regardless of which face of the block rests on the table, but that force is “smeared” over a larger area for a larger face and over

a smaller area for a smaller face. Thus, if we divide the total force W (i.e., the weight of the block) by the area A of the given face, we obtain different values for different faces, with the largest value applicable to the face of smallest area. To this kind of number, a force per unit area, we give the name “pressure.” We say that the block exerts a pressure on the table and that the table exerts a pressure on the block. In SI units, pressure would be expressed in N/m^2 or, in our still widely used non-metric system, in lb/in^2 . (In the SI system, 1 N/m^2 is given the name 1 Pascal, abbreviated 1 Pa.)

Note that pressure is *not* simply a force, period; it is a very different, although related, concept. Although the total upward force on the block is the same regardless of which face of the block rests on the table, the pressure is quite different for each face, being largest under the face with the smallest area. In accordance with Newton’s third law, since the table exerts an upward force on the block, the block exerts a downward force on the table, and any object placed between the block and the table will be squeezed between the two. Thus, if we were to put a rubber pad between the block and the table, the pad would be squeezed down into a thinner layer under a smaller face, owing to the larger pressure, than it would be under a larger face, even though the total force W applied to each surface of the pad would be exactly the same in each instance. If we were to taper one side of the block down to a very thin knife edge and rest the block on this edge, the pressure along the knife edge would become extremely high, and, under such circumstances, the table surface might well be damaged. (Translate this discussion into visualizing the effects on a floor of pointed heels or of spikes on athletic shoes.)

When we speak of “atmospheric pressure” at the surface of the earth, we are talking about the force per unit area (N/m^2) at the base of the entire column of air that rests on the earth’s surface. When we speak of the pressure at the bottom of the ocean, we are talking about the force per unit area at the base of the column of water with the column of air on top. In each of these instances, we would calculate the pressure exactly as we did in the case of the block on the table: The total weight of the column divided by the area of its base.

Question 2.4.1 Suppose you push a block against a wall with a force of 75 N. The block surface in contact with the wall has dimensions of 25 x 37 cm. Calculate the pressure between the block and the wall. [Ans. 810 N/m^2 or 810 Pa]

Question 2.4.2 Estimate the pressure at the soles of your feet when you are standing still. Give the result both in Pa and in lb/in^2 .

Question 2.4.3 Let us examine in some detail the physics of pressure when a book lies on the table. There is more going on here than initially meets the eye.

- (a) Estimate the pressure underneath your textbook when it rests on the table. (You will have to estimate both the weight of the book and the relevant area.)

- (b) What is the downward force exerted by the atmosphere on the upper surface of your textbook when it rests on the table? (Take normal atmospheric pressure to be about 15 lb/in^2 or about $1.0 \times 10^5 \text{ Pa}$.)
- (c) Given this enormous downward force exerted by the air, how do you account for the fact that you can lift your book with so little trouble? (This question is not a trivially simple one. It anticipates ideas that will be developed in the next section. Deal with it as best you can at this point, and prepare to come back to it later. It is important to realize that air is not absent from the region between the book and the table; if it *were* absent, you would not be able to lift the book.)

2.5 PRESSURE IN FLUIDS

Fluids (i.e., liquids and gases) behave very differently from solids when they are squeezed, and the difference in behavior gives us a deeper insight into the nature and significance of the concept of “pressure.” Let us visualize some of the differences. Suppose we pour a puddle of water onto one region of the surface of a table and place a piece of plywood on another region. We take two solid rectangular blocks, such as the one in Fig. 2.4.1, and place one on the water puddle and the other on the piece of plywood as shown in Fig. 2.5.1. Visualize the difference in behavior: The water does not stay under the block; it squirts out to the sides until only a very thin film is left under the block. (Recognize, in passing, that exactly the same thing must happen to the air initially between the block and the table when you put the block on the table.) The sheet of plywood, however, does *not* squirt out to the sides as does the water. It bulges sideways very, very slightly as it is compressed and deformed by a very heavy block, but it remains under the block with only a very small decrease in thickness.

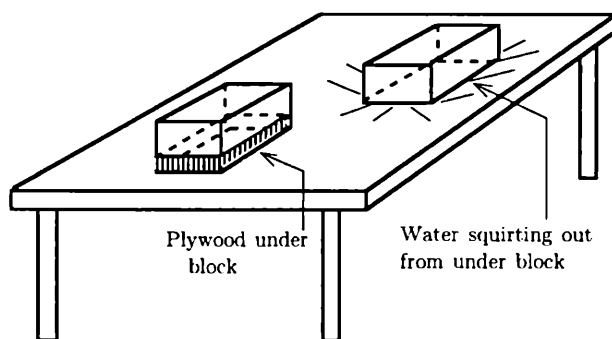


Figure 2.5.1 One block placed on a sheet of plywood and the other on a layer of water.

By squirting out to the sides (i.e., accelerating horizontally), the water is indicating that it is being subjected to an unbalanced force in the horizontal

direction. How can the water be subjected to an unbalanced force in the horizontal direction when the force exerted by the table on the block and the force exerted by the block on the table are both in the vertical direction? And what is the difference between the layer of water and the layer of plywood, i.e., how is it that the solid plywood is *not* subjected to an unbalanced horizontal force making it squirt out to the sides?

To help visualize what is involved, let us perform a “thought experiment” as follows: Imagine a cylinder containing a quantity of water as in Fig. 2.5.2. Into this outer cylinder we slide a closely fitting solid cylinder which acts as a piston resting on the water after we let out any trapped air. (The piston fits so neatly into the outer cylinder that no water leaks out between them.)

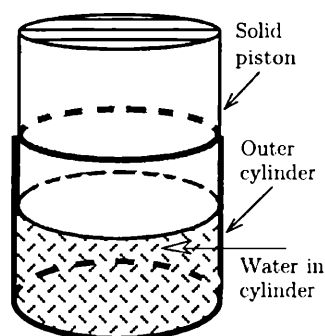


Figure 2.5.2 Solid piston resting on water in a cylinder with no leakage between the piston and the cylinder.

The water exerts a uniformly distributed upward force on the piston, and the total upward force exerted by the water must be exactly the same as the total upward force that would be exerted by a table if the piston were resting on the table, i.e., the upward force must be equal to W , the weight of the piston. The pressure p at the piston-water interface must therefore be equal to W/A , where A denotes the area of the base of the piston (or the cross sectional area of the cylinder, since the two areas are essentially equal.)

Now what happens if we puncture a hole in the wall of the cylinder somewhere below the bottom of the piston? Water will, of course, squirt sideways out of the hole (as it does when we simply put the piston on a water puddle on the table surface) regardless of where we punch hole around the periphery of the cylinder. Any chunk of water located at the opening is apparently being subjected to an unbalanced force in the horizontal direction and is accelerated outward. We therefore also infer that chunks of water at the cylinder wall, not at an opening, must be subjected to balanced horizontal forces since these chunks are not being accelerated, i.e., the chunk of water exerts an outward force on the cylinder wall, and the cylinder wall exerts an inward force on the chunk of water.

If we place a solid material in the cylinder and rest the piston on it, the solid does *not* tend to squirt out of a hole in the cylinder. In other words, with a solid material the force in the vertical direction is *not* accompanied

by a net accelerating force in the horizontal direction on chunks of the solid. In the case of the liquid, however, the vertical force exerted by the piston is accompanied by an accelerating force in the *horizontal* direction on chunks of the liquid.

If we want to confine the water at the hole (by closing the hole with our finger, for example), we must exert a force in the horizontal direction opposing the exit of the water. The basic factual questions that arise are: What is the size of the force that we must exert? How does it depend on the size of the hole and on the weight W of the piston? The questions can be answered by making direct measurements, and the experiments yield a remarkably simple and highly significant answer: The total force we must exert over a hole of area A_H right at the bottom of the piston is equal to pA_H . In other words, the sideways pressure exerted by the water on the cylinder wall at the level of the piston-water interface is exactly the same as the pressure $p = W/A$ in the vertical direction at the interface. This restricted observation suggests, but does not prove, the following generalization: The pressure at any given point in a stationary fluid is the same in any direction we choose to examine at that point, not just horizontally and vertically. This generalization turns out to be correct and is known as “Pascal’s law” [for Blaise Pascal (1623-1662), the French philosopher and mathematician who first articulated this insight on the basis of essentially theoretical arguments.] To emphasize the physical idea we have developed, the pressure obeying Pascal’s law in a stationary fluid is usually referred to as “hydrostatic pressure.”

Let us immediately repeat and emphasize, however, that Pascal’s law applies *only* in fluids and not in solids. How does the situation in solids differ? The key lies in the fact that solids can resist the effect we call “shear” (Fig. 2.5.3) while fluids do not. (You can easily simulate the effect we are describing if you take your book and subject it to forces such as those illustrated in Fig. 2.5.3 (b). The deformation is simply invisibly small in less deformable solids, but it is nevertheless present.)

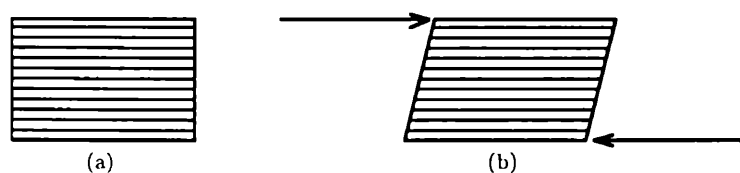


Figure 2.5.3 (a) Layers of material within a solid object not subject to external forces. (b) Layers of material within a solid body subject to forces in the direction indicated are displaced laterally. The deformation shown is called “shear.”

In solids the displacement shown in Fig. 2.5.3 (b) increases with increase in the applied force, and the solid springs back to its original shape when the

forces are removed. If the forces are sufficiently high, the solid may become permanently deformed and does not spring back. In either case, however, the material does not simply flow freely as does a fluid. The free flow of the fluid prevents the building up, under static conditions, of significant shear forces such as those shown in Fig. 2.5.3 (b).

It turns out that small shear forces such as those shown in the figure can be present in a fluid as long as *motion* continues, e.g., in a layer of lubricating fluid between a rotating shaft and a stationary bearing. This effect arises through the effect of fluid friction called “viscosity” and is present only so long as the fluid is in motion, layer over layer. If a shearing effect is applied to a fluid under static conditions, the fluid is displaced until all motion ceases as the forces drop to zero. In other words, fluids can only be compressed or expanded in all three dimensions simultaneously. If you try to subject them to shear, they will flow into whatever shape eliminates the attempted shearing effect. This how a fluid always manages to fill (and take the shape of) the container into which it is poured.

Now let us visualize what must be taking place in a tall cylinder of fluid subjected to the effect of the earth’s gravity. If we draw an imaginary horizontal plane through the column at any given level, we realize that the fluid below this plane must be exerting an upward force on the fluid above the plane and that this force must be equal and opposite to the weight of the upper fluid. Therefore the *pressure* at any level of the column must increase as we go down or decrease as we go up, and thus gravity plays an important role in determining the pressure at any level in a fluid column of significant height. This is how the ambient (surrounding) pressure becomes exceedingly high at the bottom of the sea. For this same reason, the pressure p near the bottom of the cylinder in Fig. 2.5.2 must be somewhat higher than the pressure at the piston-water interface. If we deal with fluid samples that have relatively small vertical height, we can take the gravitational effect to be negligible, and a single value of the pressure p applies to the entire sample. This would be the case, for example, in a sample of gas or liquid in a container in the laboratory.

Now let us examine the physical significance of the fact that we live at the bottom of the sea of fluid that we call “air.” What we call “atmospheric pressure” and measure with barometers is the pressure at the bottom of this sea. Each square inch or square meter of surface at the ground is supporting the weight of the corresponding column of air that extends from the ground to the upper reaches of the atmosphere. At any point in the air around us, the pressure is the same in all directions, up, down, sideways, and at any other angle, in accordance with Pascal’s law.

Question 2.5.1 Return to Question 2.4.3 regarding the fact that one can lift one’s book off the table so easily despite the enormous downward force exerted by the atmosphere on the top of the book. In the light of our discussion of hydrostatic pressure and the fact that it is the same in all directions in the fluid, how do you account for the fact that the pressure of the atmosphere does not inhibit

your lifting the book? (Keep in mind the fact that you never eliminate the film or layer of air between the book and the table, even though it becomes very thin, nor do you sever the connection between this layer and the surrounding air.) What would the situation be like if you were able to eliminate the layer of air between the book and the table?

Question 2.5.2 In the light of what we have said about living at the bottom of the sea of air, how do you account for changes with time in barometer readings at a fixed observing station at the ground, i.e., for variations in atmospheric pressure at the ground? (Note: This is an open ended question concerning a very complex physical problem. You are being invited to visualize the possibilities and the phenomena. Do not expect a rigid, pat answer. Keep in mind, however, that the air is continually in motion despite the existence of frictional effects.)

Question 2.5.3 Let us consider the pressure on the bottom of each of the two containers of water shown in Fig. 2.5.4. Container (a) is a simple circular cylinder with cross sectional area denoted by A_1 . It is filled with water to a height h , and the total weight of water is denoted by W_a . Container (b) consists of a narrower cylinder with cross sectional area A_2 fitted into a larger one with cross sectional area A_1 , the same as that of container (a). There is a solid cover at the top of the larger section. The height of the large lower section is h_1 , and the height of water in the upper section is h_2 . The total height h of the water is the same in both containers (i.e., $h = h_1 + h_2$). The pressure on the bottom of container (a), over and above the pressure that would be exerted by the atmosphere, must be given by W_a/A_1 .

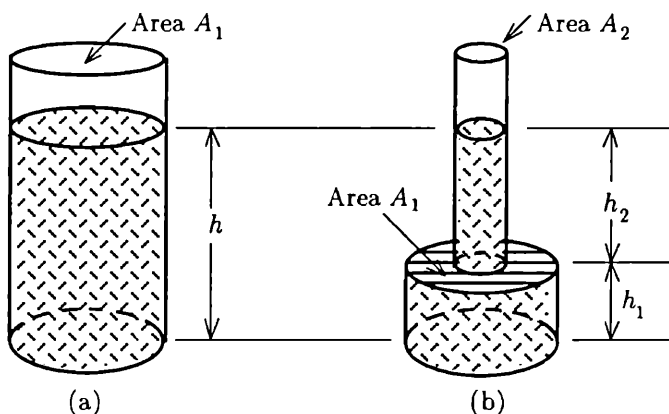


Figure 2.5.4 Cylinders containing water. The total height of the water column is the same in both (a) and (b), i.e., $h = h_1 + h_2$.

- (a) We now raise the following question: How do the pressures on the bottom of the two containers compare, i.e., are the pressures equal or is one greater than the other?
- (b) In response to this question, a person makes the following argument: Since the weight of water in container (b) is obviously less than that in container

- (a), and since the areas at the bottom are the same, the pressure must be lower at the bottom of (b) than at the bottom of (a).
- (c) How do you respond to this argument? (Answer: The fact is that the pressures at the bottom are the *same* in both containers, the pressure in (b) being determined by the total height of the central column and not by the total weight W_b of water in the container. Explain how this comes about by drawing force diagrams of different columns of water in (b) and by making use of Pascal's law. Note, for example, that, in the light of Pascal's law, columns of water outside the central column in (b) must interact with the upper cover, and, as a result, the pressure at the bottom of these columns is not simply determined by the weight of the column alone.)
- (d) Describe how the pressure varies as we start at the bottom of each container (a) and (b) and go upwards to the water surface.

Now let us return briefly to the role of pressure in determining the state of a system when the height is small, and gravitational effects are negligible: In such cases, the pressure must be uniform *throughout* the system, and a single value of p is sufficient to describe that aspect of the state.

Question 2.5.4 Suppose we start with a container of gas or liquid in which gravitational effects are negligible but the pressure is higher at one end of the container than at the other. Describe what will happen in the container under such circumstances. When will movement of fluid cease?

2.6 THERMOMETERS AND TEMPERATURE

Since we have been relying on your familiarity with ordinary thermometers and their temperature readings, we have mentioned the term “temperature” quite frequently in the preceding sections without trying to define it more precisely. That is where we shall leave matters for the time being. When we refer to “temperature,” we mean the reading on an ordinary thermometer. (It turns out that temperature, as a basic physical concept, can eventually be defined in terms of energy concepts and without reference to the special properties of any substance whatsoever. This is done through the second law of thermodynamics, an insight that you will encounter further along in your study of physics and one that was not attained until near the middle of the 19th century—long after thermometers came into common use and contributed to the insights we discuss in this and following sections.)

All thermometry is based on observing changes in some “thermometric property” of a material as it becomes hotter or colder. Familiar alcohol and mercury thermometers utilize the expansion and contraction of a liquid confined under constant pressure in a glass tube. The constant-volume gas thermometer utilizes increase and decrease in pressure of a gas held at constant

volume in a rigid container. Other thermometers are based on changes in electrical resistance of a metal (e.g. platinum) or some other material; changes in electrical potential difference at a contact between two different metals (thermocouple); changes in the spectrum of light emitted by a glowing substance (at very high temperatures); and so forth.

Galileo invented a crude "thermoscope" (the suffix "scope" implies qualitative indication without numerical calibration) around 1600. It consisted of a glass bulb containing air and having a long, narrow stem which extended downward into a vessel of water. As the bulb became hotter or colder, the air within it expanded or contracted, and the water level rose or fell in the stem. This device was, of course, sensitive to atmospheric pressure changes as well as to higher or lower temperature.

Various improved thermoscopes were made by subsequent investigators. Sensitivity to pressure variations was eliminated by use of liquids completely sealed in complicated glass bulbs or tubes. Toward the end of the 17th century, quantitative measurement was introduced through the calibration of thermometer scales by marking fixed points (such as the melting of snow and melting of butter) and dividing the scale between the points into some arbitrarily chosen number of uniform intervals. Newton performed experiments with such a device.

Between the years 1714 and 1717, D. G. Fahrenheit (1686-1736) constructed reliable alcohol and mercury thermometers using simple cylindrical glass tubes and bulbs and proposed the scale that still bears his name: Freezing point of water (called the "ice point") marked at 32° and boiling point of water at standard atmospheric pressure (called the "steam point") marked at 212° , with 180 uniform divisions between the two arbitrary points. During the period between 1710 and 1743 there also evolved the centigrade, now called the "Celsius," scale and named after the Swedish astronomer Anders Celsius, who, during the 1740s, strongly advocated a centesimal scale (100 uniform divisions) between the ice point and the steam point.

2.7 THERMAL EQUILIBRIUM

The development of reliable calibrated thermometers made possible the quantitative investigation of phenomena of heating and cooling. (Such investigations were being carried on during the same years that "electricians," as they called themselves, were clarifying the phenomena of electrostatics.) Among the principal investigators of thermal phenomena, was the Scottish physician Joseph Black (1728-1799) who was, at various times, professor of medicine and chemistry at the Universities of Glasgow and Edinburgh.

Given the ability to make quantitative observations, Black made a systematic investigation of changes in state associated with what, in Section 2.1, we called "thermal interaction." He first arrived at the conclusion that, when materials at different temperatures are brought together in thermal interaction

(while being “insulated” from interaction with other objects or the surroundings), they all end up at a state characterized by the same final temperature. Generalizing the outcome of this class of observations, he wrote:

By the use of thermometers, we have learned that, if we take . . . different kinds of matter—such as metal, stones, wood, cork, feathers, wool, water, and a variety of other fluids—although they be all at first at different temperatures, and if we put them together in a room without fire and into which the sun does not shine, [they will undergo thermal interaction such that the temperature of the hotter bodies will decrease while that of the colder bodies will increase] during some hours perhaps, or the course of a day, at the end of which time, if we apply a thermometer to them all in succession, it will give precisely the same reading. [The bodies are] thus brought into a state of equilibrium. . . . We must therefore adopt, as one of the most general laws of [thermal interaction], the principle that all bodies communicating freely with one another and exposed to no inequality of external action, acquire the same temperature as indicated by a thermometer. [We have, in the preceding quotation, placed brackets around passages in which we have replaced Black’s older terminology with our modern forms.]

Although it is readily apparent in everyday experience that, when bodies at different temperatures are brought into thermal contact, the temperature of the higher temperature bodies invariably decreases while that of the lower temperature bodies increases, it is far from obvious that the bodies all end up at the same temperature. Different materials such as stone, metal, wood, or cloth *feel* entirely different to the touch when they are at exactly the same temperature, and Black’s insight could only have been attained through quantitative measurement and carefully constructed experiments.

Black’s generalization concerning thermal equilibrium incorporates an idea that is stated more analytically in modern textbooks under the name of the “zeroth law of thermodynamics”: If Body A is in thermal equilibrium with body B, and B is in equilibrium with body C, then A and C will also be found to be in thermal equilibrium.

Question 2.7.1 Suppose that an object (insulated from its surroundings) has a higher temperature at one end than at the other. What do you expect will happen to this non-uniform temperature distribution as time goes by? Why is it that only one single value of temperature is needed to describe the state of a system in thermal equilibrium? How do you account for the fact that the earth, as a whole, is *not* in a state of thermal equilibrium?

2.8 THE CONCEPT OF “CONSERVATION”

In Chapter 1 we established that linear momentum (the vector quantity $m\vec{v}$) was conserved in closed systems of mechanically interacting objects. In other words, the final total momentum of the system was the same as the total initial momentum of the system regardless of the momentum changes imparted to individual objects within the system. Are any other properties of natural systems conserved?

Around the middle of the first century BC, the Roman poet Lucretius wrote in his philosophical poem *De Rerum Natura* (“Of the Things of Nature”):

[Superstition] cannot be dispelled by sunbeams, the shining shafts of day, but only by an understanding of the outward forms and inner workings of nature. In developing this theme, our starting point will be this principle: Nothing can be created by divine power out of nothing. . . . If things were made out of nothing, any species could spring from any source and nothing would require seed. Men could arise from the sea and scaly fish from the earth, and birds could be hatched out of the sky. . . . The second great principle is this: Nature resolves everything into its component atoms and never reduces anything to nothing. If anything were perishable in all its parts, anything might perish all of a sudden and vanish from sight. . . .

Lucretius does not use the word “conservation,” but this is a statement about conservation of material substance. The theme of conservation of some entity in natural phenomena—a theme frequently adopted on metaphysical premises—sounds and resounds in writings of the earliest natural philosophers. To Descartes it was inconceivable that the absolute numerical magnitude of “quantity of motion” (momentum) could fail to be conserved, and he deduced erroneous conclusions from this erroneous premise. (Remember that we found vector momentum to be conserved, not its magnitude.)

Leibniz convinced himself that the quantity mv^2 (called vis viva or “living force” in his day) and a product of force and displacement must be conserved, and, although not completely correct, he did attain a preliminary insight into what we now call a restricted principle of conservation of mechanical energy. Huygens asserted a principle of impossibility of perpetual motion in articulating his conviction that one cannot extract something from nothing in harnessing natural phenomena to produce useful effects.

During the second half of the 18th and first half of the 19th centuries, these vague intuitions—based on a sense of orderly connection between initial and final conditions in processes of physical change—came to be expressed as scientific laws or principles. These principles, which grasp some characteristic permanence in what would otherwise be a chaotic flux of events, are among the most powerful generalizations available to science. Not only do they unify

ranges of phenomena that would otherwise seem completely disparate and unconnected, they provide the very basis for what we refer to as “understanding” and “explanation” in scientific thought. One of these governing principles, the law of conservation of momentum, we became acquainted with in Chapter 1. We shall now examine the next principle that was established in the historical sequence, that of conservation of mass.

2.9 LAVOISIER AND THE LAW OF CONSERVATION OF MASS

Profound alterations in the appearance and character of material substances are obvious in processes such as combustion, refining of ores, actions of acids and alkalis, and in the many other similar changes to which we give the name “chemical.” We recognize that chemical processes involve not only readily apparent changes in the appearance, properties, and masses of liquids and solids, but are also frequently accompanied by evolution or absorption of gases. Furthermore, we recognize profound differences between different gases—differences residing in properties such as density, toxicity, and chemical interactions with other materials.

These realizations, although obvious to us through long established vocabularies of description and through early education, were not easily won by 17th and 18th century investigators. Many techniques of observation and measurement had to be invented and developed before it was realized that all gases were not air (even when they were colorless and tasteless). It took many years of accumulated observation to show that gases could be distinguished from each other by properties such as the ability to support or stifle combustion, produce toxic or non-toxic effects on living organisms, or react differently in contact with other substances. The famous instance of the discovery of oxygen through the separate investigations of Scheele, Priestley, and Lavoisier was really not so much a sharply defined discovery as a slowly won awareness of the significance of experimentally observable differences between this gas and air.² It turned out, however, that this slowly won awareness played a major role in breaking through to the next great conservation law.

The second half of the 18th century saw the development of very accurate weighing techniques (gravimetric methods) and greatly improved methods of containing and handling gases. With the development of such techniques, researchers began to explore the changes in weight to be measured in chemical and thermal phenomena. Joseph Black, in the course of his studies of heat in the mid 18th century, performed experiments in which he accounted for the

²For interesting and readable analyses of this important episode in the history of science, see Chapter 6 of *The Edge of Objectivity* by C. C. Gillispie, Princeton University Press, 1960 and Chapter 6 of *The Structure of Scientific Revolutions* by T. S. Kuhn, University of Chicago Press, 1960.

weights of all materials in certain chemical reactions, but the full import of these observations was apparently not clear to Black himself or to his contemporaries. Clarification of the fundamental principles involved came through the work of Antoine Laurent Lavoisier (1743-1794) and his attack on the then prevalent theory of chemical change.

At the middle of the 18th century, the most widely held theory of chemical change assumed the existence of an imponderable (weightless) fluid called "phlogiston," which was taken to be the essence of flame or fire. When ores heated with charcoal formed metals, it was visualized that phlogiston from the charcoal combined with the ore as a "metallizing principle." When charcoal burned in air, its phlogiston escaped and combined with the air. Burning ceased in an enclosed space when the air was fully saturated with phlogiston.

Lavoisier undertook the careful gravimetric study of chemical changes in closed systems, making sure that no material (in gaseous or any other form) was allowed to enter or escape during the changes being studied. He showed that metals such as mercury, when heated gently in the presence of air, form "ores" that weigh *more* than the original metal. He further showed that some of the air, initially enclosed with the metal, disappears in the process. He showed that, when the ore thus formed was heated at a much higher temperature, it was decomposed to restore the original amount of metal and the original amount of air. He recognized in the course of such investigations that air consisted of two different gases, and he showed that one of these (the one now called "oxygen" and that Priestley had called "dephlogisticated air") combined with the metals, or was restored on decomposition, in the changes he was studying. These discoveries made the phlogiston theory untenable, and it was rapidly abandoned during the 1780s following publication of Lavoisier's research.³ (During these same years, Count Rumford, whose work we shall be discussing in more detail later, was utilizing his very sensitive balance to show that objects did not change their weight measurably during substantial changes in temperature.)

Although it had been quite apparent that the total mass of closed systems of interacting bodies did not change in the course of mechanical interactions, it was far from clear as to what happened in the vast range of other interactions known to take place. Lavoisier's careful gravimetric experiments led not only to an understanding of the nature of what we now call "oxidation" and to the downfall of the phlogiston theory but also to a clear enunciation of the principle of conservation of mass. Summarizing, in his textbook of 1789, the results of numerous painstaking experiments, Lavoisier wrote:

We must lay it down as an incontestable axiom that, in all operations of art and nature, nothing is created; an equal quantity of matter exists before and after the experiment . . . and nothing

³For more detailed description of the overthrow of the phlogiston theory, see Case 2 in *Harvard Case Histories in Experimental Science*, Harvard University Press, 1957

*takes place beyond changes and modifications of these [materials.]
Upon this principle the whole art of performing chemical experi-
ments depends.*

Note that in this statement Lavoisier implicitly connects mass (which is what he was actually measuring) with “quantity of matter.” Although some writers still make this connection, many, who are more critical of concepts and terminology, do not. Although “mass” is operationally defined in dynamics, the term “quantity of matter” has never been clearly defined. It is a loose term that has its uses, but it should not be taken as a clear and unambiguous physical concept. We shall confine ourselves to speaking of conservation of mass rather than of quantity of matter.

Question 2.9.1 In terms of conservation of total mass, describe in qualitative detail what happens in each of the following changes. Is volume conserved in these changes? Explain your reasoning. (Be sure to define clearly what you take as the system to be discussed.)

- (a) A stone is broken up into smaller pieces by hammering. Take the system to be the stone and its fragments, not including the surrounding air. Is total surface area of exposed stone conserved? Explain your reasoning.
- (b) A sheet of paper burns in a closed container filled with air, leaving a small heap of ash and producing carbon dioxide and water vapor.
- (c) A lump of metallic zinc is dropped into a beaker containing hydrochloric acid. The zinc disappears, and hydrogen gas is evolved.
- (d) A lump of sugar is dropped into a beaker of water. After stirring, the sugar has disappeared.
- (e) The temperature of a metal rod is elevated from 20°C to 90°C , and the rod expands.

2.10 QUESTIONS AND PROBLEMS

2.10.1 Give an operational definition of the term “density,” i.e., describe in your own words the measurements you would make on a chunk of material and the calculations you would make with the results of your measurements in order to obtain the number to which we give the name “density” of the material.

2.10.2 Taking Pascal’s law as a starting point, explain why a column of water will not remain vertical but will slump down into a puddle in the absence of a surrounding container while a column of solid material does not require a container to remain standing.

2.10.3 Suppose we have a vertical column of height h and cross-sectional area A of material that is homogeneous throughout (“homogeneous” means that the material is *uniform* in its properties) with a uniform density denoted by ρ . The material might

be a fluid (liquid or gas) held in a container, or it might be a solid that does not require a container to maintain a vertical column.

In the following, we shall concern ourselves with a value of pressure, denoted by p_g , over and above (i.e., in excess of) atmospheric pressure, denoted by p_{atm} . p_g is given the name “gauge pressure” because it is the value measured by ordinary gauges (such as tire gauges) that start at atmospheric pressure as a zero level. That is why we have chosen the “g” subscript. A total pressure, including atmospheric, is called “absolute pressure,” and we shall denote it by the symbol p_{abs} . Thus $p_{\text{abs}} = p_{\text{atm}} + p_g$.

- (a) Argue that the gauge pressure p_g at the base of the column is given by $p_g = \rho gh$, and that the absolute pressure would be given by $p_{\text{abs}} = p_{\text{atm}} + \rho gh$. Explain your reasoning completely and carefully with the help of relevant force diagrams.
- (b) Since pressure is defined as “force per unit area,” one might expect the formulas for the pressure to contain the area A . Explain why A is *not* present in the equations that have been derived.
- (c) How does the pressure against the wall of container vary as you go down from the surface of the liquid, level by level to the bottom? Explain your reasoning with the help of relevant force diagrams.
- (d) Suppose the density of the material is not constant and varies with the height h . What information would you have to have in order to calculate p_g , and how would you make use of this information?

Note to the instructor. Additional questions and problems relevant to this chapter can be found in Chapter 12 of Part II. Questions 12.1 - 12.9, 12.27, and 12.28 are especially suitable if adapted to the prior background that students have acquired at this point.

Chapter 3

The Concept of “Heat”

3.1 DISTINCTION BETWEEN “HEAT” AND “TEMPERATURE”

So far in this discussion, we have carefully avoided using the word “heat.” The reason for this is the following: Prior to Black’s time, the words “heat” and “temperature” were used more or less interchangeably (and many people still tend to do so to this day.) Fahrenheit, for example, speaks of “degrees of heat” when reporting temperature measurements. Black made the very significant conceptual refinement of separating the two concepts and giving the term “heat” its modern technical meaning. Let us see why such refinement is needed and what the term “heat” means (and does *not* mean) in present day science.

As indicated in Sect. 2.6, we take the word “temperature” to stand for the number observed on the scale of a thermometer. We shall reserve the word “heat” for a new idea: The interaction we perceive to be associated with temperature changes. This interaction is not the same thing as the readings on a thermometer. (The following story illustrates the fact that scientific concepts are invented by acts of human imagination and intelligence and are not like objects having a separate existence, being “discovered” by someone stumbling upon them.)

Let us start by visualizing a series of familiar thermal interactions and noting that thermometer readings alone do not account for everything that happens in the interactions. There are profound differences between various interactions in which the thermometer readings are all exactly the same.

Case 1: Consider two identical quantities of water each at exactly the same initial high temperature. We allow each sample to interact thermally with the air in the room (not with each other.) The situations, however, are not identical. Sample (a) is contained in a metal sauce pan while sample (b) is contained in a thermos bottle (or in a pan that is wrapped in thick material of the kind used in insulation of houses.) We thus have two different systems (the

water samples and their respective containers) each interacting with the same system, namely the room. We know from everyday experience that sample (a) reaches the temperature of the room long before sample (b) (that is why the thermos bottle was invented!) The point for this discussion, however, is that, although the initial and final temperature readings for the two samples are exactly the same, the *times* taken for the two temperature changes to occur are very different. The two interactions with the room differ profoundly, although we would not know this if we were only told the initial and final temperature readings of each sample. The thermometer readings *alone* are certainly not telling us the whole story of the interactions that are taking place.

Question 3.1.1 In the light of the preceding discussion, describe why it is desirable to provide thermal insulation in houses. In what way is the house analogous to the thermos bottle?

Case 2: Consider two different quantities of water (one large and one small) each in a similar container placed over identical gas burners or identical electrical heating elements. We thus have two different systems (the two quantities of water), each interacting with an identical system (the burners.) Suppose we start heating the two containers at the same time at the same initial temperature of 20°C and raise the temperatures to 95°C . The initial and final temperatures of both samples are exactly the same.

Question 3.1.2 What would you miss knowing if were told only the initial and final temperature readings? Describe the *differences* between what you know happens in the two cases—differences in the interactions that are *not* revealed by the thermometer readings alone. (Note that time difference is *one* element, as in Case 1, but that there is more to it than just a difference in time intervals. What about the *costs* involved?)

Case 3: Consider a beaker containing crushed ice that has just been taken out of a refrigerator at a temperature of -2.0°C and placed in a room where the air temperature is 20°C . We are interested in the thermal interaction between the system consisting of the beaker of ice and the system that is the room. We keep stirring the contents of the beaker so as to keep the temperature in this system as uniform as possible. The temperature in the beaker rises to the ice point (0°C) and then stays at that temperature as liquid water forms through melting of the ice. The temperature in the beaker begins to rise toward room temperature only after all the ice has melted.

Question 3.1.3 How do we know that thermal interaction with the room is continuing even during the period when the temperature in the beaker is not changing at all, i.e., what is the evidence for continuing thermal interaction even though the thermometer reading in the beaker remains constant? (Hint: What do we

see happening to the ice? What would happen if the room temperature, instead of being *above* 0°C , were at exactly 0°C ? below 0°C ?) Does constancy of temperature of a system necessarily mean that no thermal interaction is taking place with another system?

Case 4: It is a matter of common experience that, when two systems at different temperatures are brought together so as to interact thermally, the temperature changes in coming to equilibrium are *not the same* in each of the two systems.

Question 3.1.4 Draw on your own personal experience to give specific illustrations of the fact that, when the two systems contain different masses of the same material (e.g., water), the temperature change of the larger mass of material is always smaller than that of the smaller mass (providing there is no thermal interaction with some third system). Under what circumstances are the temperature changes equal? What happens to the temperature change of the larger system as its mass increases without limit while the mass and initial temperature of the smaller system remain the same?

The illustrations assembled in Cases 1 through 4 all testify to the idea that, in thermal interactions, there seems to be a *process* involved in which one system influences the state of the other and that thermometer readings alone do not fully describe the interaction. Sometimes they do not even reveal that an interaction is taking place. To this process of interaction that we now begin to recognize, we give the name “transfer of heat,” and Black is to be credited with redefining the term “heat” in this sense, distinguishing it unambiguously from “temperature,” and thus introducing our modern scientific terminology in which we must not use the words “heat” and “temperature” interchangeably. It should be clearly noted that, when we speak of transfer of heat (or, for convenience, shorten this longer form to simply the words “heat,” or “heating”) we are not speaking of a state variable in the sense that velocity, temperature, pressure, density, or concentration are state variables; we are speaking of a *process* that alters the state variable temperature or causes melting or freezing (referred to as “change of phase”).

We say that heat is gained by, or transferred to, systems when their temperature increases (or when they melt) in thermal interactions, and we say that heat is lost by, or transferred out of, systems that decrease in temperature (or freeze).

Let us note explicitly exactly what we are *not* doing: We are *not* identifying what heat “is”; we are *not* describing a mechanism of interaction; we are proposing *no* model or theory for the nature of “heat”; we are *not* visualizing what might be happening in such interactions on the microscopic scale of atoms and molecules. We are simply giving the name “transfer of heat” to a

process we discern to be taking place without knowing anything else about it. Models for the nature of heat [based on the notion that it was a conserved, imponderable (i.e., massless) fluid] were proposed and used during the 18th, and into the 19th, century but were eventually found to be inadequate and were discarded. We shall encounter some of the history of these insights in Chapter 4.

Question 3.1.5 What do we do to a system to make it gain heat by thermal interaction? What do we do to make it lose heat? Give several specific examples out of everyday experience in each instance.

Question 3.1.6 Does thermal interaction take place and is heat transferred when you bring two systems at the same temperature into thermal contact? Explain your reasoning.

3.2 MEASURING AMOUNTS OF HEAT TRANSFERRED

Having invented the qualitative concept of “transfer of heat,” we are now faced with the question as whether we can make it quantitative, i.e., can we devise a way of measuring it numerically and talking about “*quantity* of heat transferred”? A reasonable basis for numerical measurement was provided by a number of Black’s predecessors and was refined and clearly interpreted by Black himself. The idea was based on observation of the behavior of water, which, because of its availability, ease of handling, and benign chemical properties, makes it an especially convenient reference material.

The basic hint comes from the following: When equal masses of water, initially at different temperatures, are mixed, it is observed that the final equilibrium temperature of the mixture is very close to half¹ way between the two initial temperatures (providing precautions are taken to insulate the system consisting of the two masses of water as much as possible from thermal interaction with surrounding air or other objects).

Question 3.2.1 Let us put this verbal statement into algebraic form using the following notation: t_H and t_L denote the initial temperatures of the higher and lower temperature water masses, respectively, and t_{EQ} denotes the final equilibrium temperature after the thermal interaction brought about by the mixing. Argue that the assertion in the preceding paragraph (to the effect that the final equilibrium temperature ends up half way between the two initial temperatures) implies that

$$t_{EQ} - t_L = -(t_{EQ} - t_H) \quad (3.2.1)$$

¹It turns out that the thermal properties of water are such that the final equilibrium temperature is not *exactly* half way between the two initial temperatures. It is very close, however, and provides the start that made possible the quantitative measurements we are describing. The concepts being generated are subsequently refined through a next level of closer approximation. Such successive approximations in developing and refining scientific concepts are very common in all scientific endeavor.

i.e., that the two temperature changes (defined, as we always define “changes”, as final minus initial values) are equal in magnitude and opposite in algebraic sign.

When *different* masses of water, initially at different temperatures, are mixed, the final temperature is *not* half way between the two initial temperatures but is always closer to the initial temperature of the larger mass. In fact, the two temperature *changes* turn out to be very nearly inversely proportional to the masses of water involved, with the larger mass undergoing the smaller temperature change.

Question 3.2.2 Letting m_H and m_L denote the masses of the higher and lower temperature quantities of water, respectively, argue that the preceding statement about temperature changes of different interacting masses implies that

$$\frac{(t_{EQ} - t_H)}{(t_{EQ} - t_L)} = -\frac{m_L}{m_H} \quad (3.2.2)$$

or that

$$m_L \Delta t_L = -m_H \Delta t_H \quad (3.2.3)$$

where Δt_H and Δt_L denote the temperature changes of the higher and lower temperature masses, respectively.

Note that the final number on the left hand side of Eq. 3.2.3 depends only on numbers applying to the lower temperature body of water while the final number on the right hand side depends only on numbers applying to the higher temperature body. Thus each side of Eq. 3.2.3 seems to describe something that “happened” to each separate body involved in the interaction. It is inviting, therefore, to think of the number $m_L \Delta t_L$ on the left hand side as representing the amount of heat *gained* (since the number is positive because of the sign of Δt_L) by the lower temperature water and the number $m_H \Delta t_H$ on the right hand side as representing the amount of heat *lost* (since the number is negative because of the sign of Δt_H) by the higher temperature water.

In terms of absolute *magnitudes*, Eq. 3.2.3 becomes

$$|m_L \Delta t_L| = |m_H \Delta t_H| \quad (3.2.4)$$

and the equality of the magnitudes of heat gained and lost (in the absence of thermal interaction with any other systems) suggests that we might be dealing with a conserved quantity for which we have now devised a method of measurement.

Question 3.2.3 Explain the preceding statement in your own words. *Why* does Eq. 3.2.4 suggest that the quantity we have invented may be conserved?

In interpreting the numbers in Eqs. 3.2.3 and 3.2.4 as quantities of heat transferred, we are immediately implying that we have also created a unit of measurement. Let us investigate what this unit must be. If we have 25.0 g of water changing temperature by 5.2 C°, the quantity $m\Delta t$ must be 130. This means that 130 times as much heat was transferred as would have been transferred in order to change the temperature of 1.0 g by 1.0 C° (i.e., for $m\Delta t = 1$). We give the name "one calorie" to the latter amount of heat—that which changes the temperature of 1.0 g of water by 1.0 C°. In changing the temperature of 25.0 g of water by 5.2 C°, we say that 130 calories were either gained or lost by the sample, depending on the direction of the temperature change.

The experimental technique of measuring quantities of heat transferred by observing masses of interacting materials and the associated temperature changes when thermal interaction takes place in a thermally isolated system is called the "method of mixtures," and a container in which the insulated experiment is conducted is called a "calorimeter." We also speak of the process of making the measurements as "calorimetry" and of the making of "calorimetric experiments."

3.3 SPECIFIC HEAT AND HEAT CAPACITY

In his experimental observations, Black showed (as might be expected) that Eq. 3.2.3 does not predict the temperature change correctly if two *different* substances are allowed to interact thermally, i.e., if one substance is water and the other is mercury, or iron, or copper, or some other material. He also showed that using volume (rather than mass) to measure the amount of material was not in accord with Eq. 3.2.3, while using mass as a measure always satisfied the relation. (Some of his contemporaries had claimed that volume was an appropriate measure.)

Black's experiments led him to the conclusion that, although other substances were different from water in their thermal behavior, each substance could be reproducibly and systematically compared with water in its capacity to change temperature on transfer of heat. For example, in the method of mixtures, it is found that 100 g of mercury always behave like 3.3 g of water, signifying that water has a much larger "capacity" for heat, in that a given quantity of heat would produce in 100 g of water a very much smaller temperature change than it does in 100 g of mercury.

Finding such comparisons to be measurable and reproducible, Black was led to describe the whole idea by simply introducing a proportionality constant multiplying the $m\Delta t$ product. Thus Eq. 3.2.3 would be modified to read

$$c_A m_A \Delta t_A = -m_W \Delta t_W \quad (3.3.1)$$

where c_A is the suggested proportionality constant.

Let us interpret each term in Eq. 3.3.1: $m_W \Delta t_W$ would represent the amount of heat gained or lost by m_W g of water in a mixtures experiment, while $c_A m_A \Delta t_A$ would represent the heat lost or gained by some other substance A. If substance A is water, $c_A = 1$. If A is some other substance, the quantity $c_A m_A$ is to be interpreted as the equivalent amount of water; e.g., in the case of mercury, for which a numerical value was cited above, 100 g of mercury were found to be equivalent to 3.3 g of water, and c_A for mercury must therefore be 0.033. Following Black, the quantity c_A is called “specific heat” or “heat capacity” of material A.

Since the product $c_A m_A \Delta t_A$ denotes a quantity of heat, c_A must have the dimensions of heat per unit mass per unit change in temperature. For mass in grams and temperature in degrees Celsius, the unit of heat is called the calorie (as indicated in the preceding section), and the units of specific heat c_A must then be cal/(g) (C°). Thus the specific heat of mercury would be given as 0.033 cal/(g) (C°), while the specific heat of water, taken as the reference substance, has been defined to be 1.000 cal/(g) (C°).

There are two other units besides the calorie with which it will be useful to be familiar. The amount of heat that elevates the temperature of 1 kg of water by 1 C° is 1000 cal and is called a “kilocalorie” or a “large calorie,” while the calorie we have defined earlier is implied to be the “small calorie.” (The “calories” that are so frequently quoted in connection with food intake and nutrition are really kilocalories, i.e., large calories, not small calories. You have to allow for the dropping of the kilo prefix when thinking about “calories” in food.) The other unit you will occasionally encounter is the amount of heat that elevates the temperature of one pound of water by one Fahrenheit degree. This is called the British thermal unit and is abbreviated Btu.

Question 3.3.1 Using Eq. 3.3.1 as a basis, describe how you would perform experiments to compile a table of specific heats, c_A , for a variety of substances. Give specific details of apparatus you would require, how you would bring the samples of water and other materials to the initial temperatures you wish to use, the procedures you would employ. Take into account the fact that the calorimeter cup or container in which the experiment is performed is part of the system; it must itself participate in the thermal interactions that take place within it even if the cup itself is well insulated from the surroundings; i.e., it is necessary, as part of the experiment, to determine the water equivalent of the container. (Black was the first investigator to recognize the need for this determination. His predecessors failed to do so.) Describe carefully the precautions you would take to minimize systematic errors. What are the possible sources of such errors? How would they affect the results of your measurement; i.e., would your calculated values tend to be higher or lower than the true ones? (After thinking through your own methods and procedures, see question 3.3.2; it may supply you with a few useful hints.)

Table 3.5.1 lists the specific heats of a number of different substances. Once such a table has been compiled through measurements such as those visualized in question 3.3.1, we can predict results that would be obtained in various thermal interactions, even those that do not necessarily involve water directly. Eq. 3.3.1 can be generalized to the form

$$c_A m_A \Delta t_A + c_B m_B \Delta t_B + c_C m_C \Delta t_C + \dots = 0 \quad (3.3.2)$$

applicable to thermal interactions in a system of different substances A, B, C, and so forth, insulated from its surroundings. Note that the quantities in Eq. 3.3.2 may be positive or negative depending on the algebraic sign of the relevant Δt . Heat is lost when Δt is negative and gained when it is positive; the gains and losses must add up to zero if heat is conserved. Experience with the concepts and calculations we have been describing, and with the fact that Eq. 3.3.2 is consistent with all observations of thermal interactions of the kind being investigated, reinforces the idea stated earlier, namely, that "heat is conserved when thermal interactions take place within a thermally isolated system." But this statement, by itself, does not provide, or even imply, a specific model or theory as to the nature of "heat."

Question 3.3.2 A calorimetry experiment by the method of mixtures has been conducted as follows: A brass calorimeter cup, having a mass of 54 g and containing 118 g of water, is insulated to some extent by being placed within another, larger cup. The still air (which is what provides some moderate insulation) surrounding the inner cup is at a temperature of about 20°C. An unknown sample of metal having a mass of 152 g is heated to 99.0°C in boiling water, quickly dried off, and placed into the water in the calorimeter cup, initially at 15.0°C. (Why the procedure of drying? What error would be introduced without it?) The final equilibrium temperature in the calorimeter cup is observed to be 24.7°C. Calculate the specific heat of the unknown material. [Answer: 0.107 cal/(g)(C°)] According to Table 3.5.1, what is probably the unknown material? Now note that the experiment was conducted in such a way that the calorimeter cup was initially at a temperature about 5° below the ambient (i.e., surrounding) temperature and ended up approximately 5° above the ambient temperature. This was not an accident. A preliminary, exploratory experiment was first conducted to ascertain the approximate size of the temperature change that would take place in the calorimeter, and the initial temperature was then deliberately set at such a level that the final temperature would end up very nearly as far above the ambient as the initial was below. What is the point of this aspect of the experimental design? What systematic source of error is being minimized?

Question 3.3.3 Consider a thermal interaction between two bodies A and B having initial temperatures t_A and t_B respectively. Suppose the mass of B is made larger and larger each time the experiment is repeated with the same initial temperatures. What happens to each of the two temperature changes as the mass of B increases? What will be the final temperature of each body in the

limit of very large mass of B relative to mass of A? Explain the connection between the analysis you have just conducted and the fact that a beaker of hot water placed on a table in a room soon ends up at room temperature without any observable rise in temperature of the room. How would you have to modify the operation in order to produce a discernible increase in temperature of the room?

3.4 REFINEMENT OF THE SPECIFIC HEAT CONCEPT

As we have seen in other instances, the initial definition of a concept frequently points the way to further investigation and experiment. Such experiments frequently lead to deeper insights which, in turn, lead to refinement and redefinition of the original concept. (Inventing the concept of average velocity, for example, leads to exploration of motion and to recognition that the concept of "average velocity" is not adequate to the description of situations in which velocity changes continuously. This perception leads to the refinement embodied in the invention of "instantaneous velocity.") Through a sequence of such successive revisions, we fashion deeper, more general, and more powerful descriptions of natural phenomena.

In the years following introduction of the specific heat concept, increasingly accurate observations over a wider variety of substances began to show that the value of c for any given substance is different depending on the physical circumstances under which the thermal interaction is allowed to take place. In particular, it was observed that gases exhibit a substantially lower value of c if the heating or cooling occurs at constant volume in a rigid container rather than in a cylinder in which a moving piston keeps the sample of gas at constant pressure. (The gas, of course, increases in volume under these circumstances.) If heating of the gas is carried out under conditions of arbitrarily changing pressure and volume, c is found to be different for each different "path" or sequence of states. Hence it becomes apparent that specific heat becomes a uniquely definable property of a material only if the *path* of the heating or cooling process is uniquely specified.

This gives us a more sophisticated view of the numbers we obtain when we measure specific heat by the method of mixtures. Such calorimetry is always conducted at constant pressure, and that defines our path. Specific heat so determined is called "specific heat at constant pressure" and is usually denoted by the symbol c_p . (The specific heats discussed in the preceding section should all have p subscripts on this account, but this subscript is usually dropped when the context under consideration is clearly and entirely that of constant pressure.)

Another readily definable path is that of the constant volume process, and specific heats so determined are denoted by c_v . It is extremely difficult to confine liquids and solids at constant volume while their temperature is increased or decreased (why?), and c_v is therefore not measured directly for condensed

materials. It turns out that c_p and c_v differ very slightly for liquids and solids (the difference for water is about 0.5%), and the difference is frequently neglected in making calculations. Gases, however, can readily be contained at constant volume, and the difference between c_p and c_v is found to be relatively large, of the order of 30 or 40%. Air, for example, has $c_p = 0.241 \text{ cal/(g) (C}^\circ\text{)}$ and $c_v = 0.172 \text{ cal/(g) (C}^\circ\text{)}$, respectively.

A second refinement induced by increasingly accurate measurement, came with the demonstration that specific heat of a substance is not constant even when the path is specified: It was discovered that c_p varies with temperature. If water is assigned the value $c_p = 1.000 \text{ cal/(g) (C}^\circ\text{)}$ (remember that, water being selected as the standard reference material, this assignment is made arbitrarily) in the range between 14.5°C and 15.5°C , it is found, with very precise measurements, that $c_p = 0.998$ near 30°C and 1.005 near 90°C . The variation is small at ordinary temperatures, but nevertheless significant and observable. If the variation had been large instead of very small, the simple calculations and the apparent conservation relation developed in the preceding sections would not have emerged so easily, and it is likely that the evolution of the heat concept would have been slower and more difficult.

Recognizing c_p to be a function of temperature [indicated by the notation $c_p(t)$], we can refine the concept by interpreting $c_p(t)$, the specific heat at a given value of t , as a quantity similar to an instantaneous velocity—an instantaneous “rate” at which heat must be supplied per gram of material per Celsius degree change in temperature. A quantity of heat Q transferred to m g of material between temperatures t_1 and t_2 must be calculated as the limit of a sum (i.e., an integral or area under a graph) just as we calculate displacement from instantaneous velocities:

$$Q = \int_{t_1}^{t_2} m_A c_{pA}(t) dt \quad (3.4.1)$$

Question 3.4.1 Explain Eq. 3.4.1 in your own words and interpret it as an area under a graph. (Sketch a graph with appropriate coordinates.) Explain in detail the analogy between this relation and the calculation of a displacement Δs when instantaneous velocity is known as a function of time (i.e., $v(t)$ is a known function). Sketch a possible graph and the relevant area for the latter case. Under what circumstances (i.e., under what relative values of initial and final temperature) does Q turn out to be positive; under what circumstances negative? Show that this is consistent with what we have said earlier about algebraic signs associated with values of Δt . Sketch the shape of the graph and the area we are implicitly talking about when we treat c_{pA} as a constant, independent of temperature. (It is, perhaps, unfortunate that the same symbol t is being used for Celsius temperature in one context and for clock readings in the other, but you will encounter such overlap occasionally and should be prepared to make the necessary discrimination when overlap arises.)

3.5 PHASE CHANGE AND LATENT HEAT

When ice melts, or a molten metal crystallizes into its solid form, or a liquid is vaporized, we say that a “change of phase” has occurred. It was clear to early investigators that phase changes had something to do with what we now call thermal interaction, but, prior to the development of a clear conceptual distinction between heat and temperature, it would have been virtually impossible to recognize the essential feature that distinguishes change of phase from ordinary heating or cooling. Black was the first to provide the necessary insight:

Melting has been universally considered as produced by the addition of a very small quantity of heat to a solid body, once it had warmed up to its melting point, and the return of the liquid to the solid state, as depending on a very small [loss] of heat. . . . It was believed that this small addition of heat during melting was needed to produce a [very] small rise in temperature as indicated by a thermometer.

The opinion I formed . . . is as follows. When ice or any other solid substance is melted, . . . a large quantity of heat enters into it . . . without making it apparently warmer when tried with a [thermometer]. . . . I affirm that this large addition of heat is the principal and most immediate cause of the liquefaction induced.

Let us note very carefully what Black was saying (and what was soon verified to be the case). It was thought that melting and freezing took place with transfer of a very small amount of heat and was accompanied by a very small increase or decrease in temperature of the material undergoing change of phase. This turns out to be incorrect. A relatively large amount of heat is transferred while the temperature of the material does not change *at all*.

Black goes on to marshal his evidence; first from the length of time it takes ice and snow to melt (recall how the element of time played a significant role in our initial progress toward distinguishing between heat and temperature in Section 3.1.)

And if the common opinion had been well founded—if the complete change of ice and snow into water required only the addition of a very small quantity of heat—the mass, though of a very considerable size, ought to be all melted within a very few minutes or seconds by the heat incessantly communicated from the [higher temperature] surrounding air. Were this really the case, the consequences of it would be dreadful . . . for even as things are at present, the melting of large amounts of snow and ice occasions violent torrents and great inundations in the cold countries. . . . But were ice and snow to melt suddenly, as they would if the former opinion of the action of heat . . . were well founded, the

torrents and inundations would be incomparably more irresistible and dreadful. . . . This sudden liquefaction does not actually happen. The masses of ice and snow require a long time to melt.

Black describes a variety of experiments designed to support his contention, among them a demonstration that the temperature of an ice-water mixture exposed in a warm room does not change as long as ice is present, while, at the same time, the temperature of an equal mass of cold water rises significantly. He argues that both the ice-water mixture and the water must be absorbing heat from the air at about the same rate.

Black then presents the results of an ingenious, simple, quantitative calorimetric experiment²:

The calorimeter consisted of a glass cup with a water equivalent of 16 cal/C°, containing 467 g of water, initially at 88°C. A piece of ice at 0°C was wiped dry, weighed quickly (404 g), and placed in the calorimeter. The equilibrium temperature was observed to be 12°C. Since the masses of ice and water are not very different from each other, one would, in the absence of the phase change, have expected the final temperature to be in the neighborhood of 40°C. The fact that the final temperature was actually very much lower than 40°C, provided strong evidence for Black's contention that a large quantity of heat must be transferred just to melt the ice at 0°C without changing the temperature until all the ice is melted. Black assumed that heat was conserved in this process just as in mixtures not involving phase change, and he calculated the amount of heat that must have been required to melt one gram of ice. We shall denote this quantity by the symbol L_f . Let us set up Black's calculation in our modern form [analogous to Eq. 3.3.2]:

$$467(1.00)(12-88) + 16(12-88) + 404L_f + 404(12-0) = 0 \quad (3.5.1)$$

Heat gained by warm water (negative quantity).	Heat gained by calorim- eter cup (negative quantity).	Heat re- quired to melt ice at 0°C.	Heat gained by water that was originally ice.
---------------------------------------------------------	-------------------------------------------------------------------	-------------------------------------------------	--------------------------------------------------------

Solving Eq. 3.5.1 for L_f , we obtain $L_f = 79$ cal/g for ice at 0°C. (Verify the calculation.)

Black showed that this was a reproducible property and introduced the name "latent heat." When melting (also called "fusion") takes place, we speak of the "latent heat of fusion." (That is the reason for the f subscript on our symbol.) Reasoning by analogy, he was sure that a latent heat was also associated with vaporization and demonstrated this to be the case. (The latent heat of vaporization of water $L_v = 539$ cal/g at 100°C.) Black ultimately measured latent heats of fusion and vaporization of other substances, demonstrating the

²Since Black reports his data in obsolete apothecaries' units, we have converted them into the units we have defined earlier in this chapter.

broad validity and significance of the concept. Table 3.5.1 lists latent heats as well as other thermal properties of a number of common substances.

Table 3.5.1 Some Thermal Properties of Various Substances					
Substance	Constant pressure heat capacity, c_p near room temperature cal/(g)(C°)	Melting point °C	Latent heat of fusion, L_f cal/g	Normal boiling point at one atm. °C	Latent heat of vaporization, L_v at normal boiling point cal/g
Air	0.241				
Alcohol (ethyl)	0.58	-117.3	24.9	78.5	204
Aluminum	0.214	660	94	1800	
Brass	0.092	940			
Bronze	0.087	1050		2300	
Copper	0.0923	1083	42	2600	1130
Glass	0.20				
Gold	0.0312	1063	15.9	2600	
Iron	0.107	1535		3000	
Lead	0.0305	327	5.47	1744	205
Mercury	0.0332	-39	2.73	357	70.6
Silver	0.0564	960	25.1	2050	560
Sodium chloride	0.210	801	124	1450	
Tin	0.0542	232	13.8	2260	
Water (liquid)	1.000	0	79.7	100	539
Water (ice)	0.49	0	79.7		
Wood	0.42				

Question 3.5.1 Examine the data in Table 3.5.1 to obtain some feeling for the order of magnitude of various thermal properties of common substances. What do you see to be special about water? What substance would you select if you wished to obtain the highest possible temperature change under transfer of a fixed amount of heat?

Question 3.5.2 A calorimeter consisting of a glass vessel (mass = 125 g) contains 1075 g of water at 10°C and is surrounded by air at about 20°C. Steam at 100°C is bubbled slowly into the water until the temperature of the system rises to 30°C. The calorimeter is then weighed, and the mass is found to have increased by 36 g. Evaluate the latent heat of condensation of the steam, setting up the

problem in the form used in Eq. 3.5.1 and describing each term in words as was done in that example.

While Black was conducting the researches we have been describing, among his students and assistants was a young man by the name of James Watt (1736 - 1819). Robinson, Black's biographer, writes that "[Watt] chanced to have in his hand, for repairs, a model of Newcomen's steam engine, belonging to the natural philosophy class, and was delighted with the opportunity which this small machine gave him for trying experiments connected with the theory of ebullition [boiling] which he had just learned from Dr. Black." Subsequently Watt's invention of the condenser and other improvements to Newcomen's engine were to play a crucial role in the advent of the industrial revolution.

Further investigation of the phenomenon of phase change shows that melting points and boiling points, as well as latent heats, change as one changes the total pressure to which the system is subjected. The freezing point of water is lowered slightly with an increase in pressure while the freezing points of many other materials increase. The effect is small for liquid-solid transitions since liquids and solids are relatively incompressible, but the effect is large for liquid-vapor transitions. For example, the boiling point of water decreases markedly and the latent heat of vaporization increases as one goes to lower atmospheric pressures at higher elevations in the atmosphere. If you go camping at high elevations, you must boil food longer than you do at ordinary elevations to get it well cooked.

3.6 HEAT IS *NOT* A FUNCTION OF STATE

In conclusion, let us return briefly to the concepts of "interaction," "system," and "state" that we developed in Sections 2.1 - 2.3. and place the concept of "heat" in that context. When we speak of "heat" and "transfer of heat," we are *not* dealing with a state variable such as pressure, temperature, density, or momentum, and, accordingly, we should *not* talk about the "heat in a system" or about the "heat *of* a system" in the manner in which we legitimately speak of the temperature or pressure or density of a system. Note that the values of temperature and pressure are unique properties or characteristics of the state of a system regardless of *how* the system might have been brought to that state. The terms "heat" and "transfer of heat," however, refer to a *process* of interaction between two systems and not to the state of either one. A process of interaction does result in changes in state variables—that is how we recognize that an interaction has taken place, and we classify the interaction on the basis of what variables have changed—but the process is not itself a state variable.

In order to calculate an amount of heat transferred, we must know the *path* or sequence in which the state of a given system changed. In our preceding

discussions, we confined ourselves to the very simple and very common special case in which the thermal interaction occurred along a path of constant pressure and in which we already knew, or in which we deduced, the constant pressure heat capacity c_p . We did not try, at this stage, to calculate heat transferred in more complicated cases along complicated paths. You will deal with such situations more rigorously in your later study of thermodynamics.

You can strengthen your insight into these concepts by now returning to the discussion of impulse and change of momentum in Section 1.9. Notice that impulse, like quantity of heat transferred, is path dependent. In order to calculate an impulse, we must know how the force delivering the impulse varied instant by instant (i.e., we must know the “path” of the force with respect to the succession of clock readings). If we have this information, we can evaluate the impulse as an integral (i.e., an area under a graph). The situation with respect to transfer of heat is exactly analogous: Delivery of impulse (which is *not* a state variable) results in a change in the state variable called “momentum.” Transfer of heat (which is *not* a state variable) at constant pressure results in change of the state variable temperature.

3.7 TEMPERATURE CHANGES WITHOUT TRANSFER OF HEAT

Change in temperature of a system is not a sure sign that thermal interaction and heat transfer have taken place. There are circumstances in which we can change the temperature of a system without subjecting it to thermal interaction with another system at a different temperature. Following are some examples:

Case 1: If we rub two solid objects together vigorously, the temperature of each object increases although no heat has been transferred through thermal interaction with a system at higher temperature. (Starting a fire by twirling a stick in contact with combustible material on a wooden block is an illustration.) Similarly, persistent stirring of a liquid increases its temperature by a small, but measurable, amount.

Case 2: If we quickly compress air in a tire pump, the temperature of the air increases although there has been no contact with a higher temperature system. If we reverse this process by starting with compressed gas in the pump cylinder at room temperature and allow the gas to expand by pushing the piston outwards, the temperature of the gas is observed to decrease although there has been no thermal interaction with a lower temperature system.

Case 3: When we turn on the switch of an element on the electric stove, the element rapidly increases in temperature although it is not in thermal interaction with a higher temperature system.

Question 3.7.1 Describe some other situations you can imagine or have actually observed in which temperature changes of objects or systems take place in the absence of thermal interaction with another system at a different temperature.

From here on, in thinking about and dealing with changes in state, we must be careful to distinguish between situations in which temperature changes are associated with transfers of heat and situations in which they are not.

3.8 QUESTIONS AND PROBLEMS

3.8.1 A chunk of copper with a mass of 500 g and a uniform temperature of 95°C is placed in a glass calorimeter cup containing 320 g of water at 10°C . The mass of the calorimeter cup is 130 g. Making use of relevant information from Table 3.5.1, predict the final temperature of the system at thermal equilibrium, indicating your line of reasoning and the idealizations and assumptions you are making. If the ambient (surrounding) temperature is 25°C , what would you expect to observe in the actual experiment, i.e., is your prediction likely to be too low or too high? Explain your reasoning.

3.8.2 How much heat must be supplied to convert one kilogram of ice at -10°C to steam at 100°C by heating at constant atmospheric pressure? Explain each step of your calculation. (Ans. 719.2 kcal.)

3.8.3 Problem 12.23 in Chapter 12 of Part II.

Chapter 4

Energy

4.1 CHANGE IN THE WORLD AROUND US

The world around us exhibits an incessant flux of change. Objects and groups of objects interact with each other and continually change their state. Positions, velocities, temperatures, pressures, compositions, electric and magnetic properties are continually being altered and, to superficial observation, the flux of change initially appears to lack order, coherence, or predictability. Yet we have seen that, as observation is sharpened and as insight develops, there turn out to be constraints operating in nature, constraints that impose certain degrees of order and indicate what can and cannot happen in various circumstances. The constraints are expressed in the two deep conservation laws we have seen so far: the law of conservation of mass and the law of conservation of momentum. The first assures us that matter will not simply disappear or appear out of nowhere unpredictably. The second indicates that, when objects exert forces on each other (as in collisions), the changes in velocity that occur are not erratic and arbitrary but are reproducible and predictable: Certain ranges of change turn out to be possible, and other ranges turn out to be impossible. In Chapter 1 we have seen cases in which we can establish connections between initial and final conditions even when we are unable to describe, step by step, the interactions that take place in between.

Let us recall briefly why we believe in these conservation laws so that we remain aware of the nature of our own thinking: As we pointed out in Sect. 1.6, this is an area of thought in which natural science is very different from mathematics. These laws are not “proved” the way a mathematical theorem is proved by deduction from a set of accepted basic postulates or principles, nor are they “derived” from some more simple starting point. After being guessed at or conjectured, the regularities these laws assert were verified, first for relatively simple systems, by direct experimental measurement and observation. They were then tested on systems other than the ones observed initially and were found to apply. In the case of momentum, the conservation law was found to be consistent with Newton’s laws of motion. We have seen

(Sect. 1.13) that the law of conservation of momentum takes precedence over Newton's third law since it applies in situations in which Newton's third law fails. (Such situations arise, for example, when separated objects, interacting electromagnetically, do not exert equal and opposite forces on each other instant by instant because a time interval, however short, elapses between the instant of change at one object and the instant of arrival of the resulting effect at the other object.)

No failure of these numerical relations has ever been found in cases where the necessary measurements can be made and the relations tested.¹ In the final analysis, we accept these statements and describe them as "laws," not because they have been "proved," but because predictions based on them have, so far, always turned out to be correct and because no one has ever observed and verified a violation. There are many systems (wind blowing over a lake, wood burning in a fireplace, a thunderstorm) that are so complex that we could never describe what happens to every parcel of material in sufficient detail to verify that the conservation laws apply. Yet, because violations have never been observed and because the predictions work, we firmly believe that mass and momentum are conserved in all phenomena, including those we are unable to analyze in detail.

Yet, as we think of various familiar phenomena, we quickly become aware that the two conservation laws alone do not tell us the whole story. Consider, for example, the simple two-body collisions we studied in developing the momentum concepts. We were forced to recognize two classes of collisions: elastic and inelastic. Mass and vector momentum are both rigorously conserved in all collisions, whether elastic or inelastic. Yet, in inelastic collisions, there seems to be an effect of "running down" of the motion which is not apparent in elastic collisions, and the conservation law does not tell us which class of collisions we are dealing with in any given instance; we must have such information in *addition* to the conservation laws we now know.

Not only does existing motion seem to keep "running down," but we begin to recognize that it seems to be impossible to produce certain effects that would, if attainable, be extremely desirable. For example, suppose that we could manufacture a ball that, when dropped from a certain height h on to solid plate, always bounced back to a height *higher* than h if not interfered with, but returned to height h if its velocity on impact was appropriately decreased. To achieve the necessary decrease in velocity, we could make the ball give a short kick to the crank of a wheel just before the bounce. The

¹We are talking here about macroscopic phenomena in the world around us. In the microscopic world of atoms, electrons, nuclei, and other subatomic particles, it has been found that the concepts of mass and momentum must both be redefined. Conservation of momentum holds for the redefined quantity, while the other conservation law applies to the combination mass-energy instead of to mass alone. All of this stems from the theory of relativity and the observations that verify the theory. These alterations of applicability at the microscopic level make no change in the application to everyday macroscopic phenomena we are considering at this point.

process would be repeated on each bounce, the ball always returning to height h . The wheel would thus keep turning, and we could use it to pump water or run an electric generator to light a light bulb. All the expense of fuel or the building of dams would be unnecessary. A sufficient number of sufficiently massive bouncing balls could keep our industrial civilization going indefinitely. That nature seems to forbid this kind of “getting something from nothing” was recognized from time immemorial, but the connection to some physical law and numerical regularity was not discerned. (No bouncing ball will, of course, ever bounce higher than the point from which it is released. The return would be to the same height in the ideal, unattainable limit in which frictional or inelastic effects are zero, and the return is always to a lower height in any real situation.)

Question 4.1.1 The device described in the preceding paragraph is called a “perpetual motion machine.” Invent one or two other such machines out of your own imagination and then describe what will actually happen, just as is done in the preceding paragraph. (You might make use of springs, swinging objects, magnets, gases or liquids in various containers, etc.)

Another indication that the laws of conservation of mass and momentum do not contain the whole story resides in the observation that certain effects and changes in state take place spontaneously in nature while others do not. Once ignited, paper burns spontaneously in contact with air, forming combustion products (mostly ash, water, and carbon dioxide) through chemical union with oxygen. Ash, water, and carbon dioxide do not “recombine” spontaneously to restore paper. When objects at two different temperatures are brought into thermal contact, the higher temperature always decreases, and the lower temperature always increases, never the other way around. (If the latter effect could be arranged to take place spontaneously under the right circumstances, we could have refrigeration at virtually no cost.) Electric batteries never recharge themselves spontaneously.

A crystal of solid salt placed in a beaker of water dissolves spontaneously in the water, the salt being eventually uniformly dispersed through the solution. A solution of salt in water never separates spontaneously into pure water and a solid crystal of salt. If we let some oxygen gas into one end of a closed container and let some nitrogen gas into the other, the gases are initially separate, but, if we wait a while, we soon find that both gases are uniformly mixed throughout the container. They have diffused or “dissolved” in each other, and they do not separate spontaneously. (We are very fortunate that this is the case. Nitrogen, being less dense than oxygen, would float on top of the latter if they separated, and, since the atmosphere consists of nitrogen and oxygen, such separation would make the world a very dangerous place in which to live.) The effects that do not occur spontaneously would not violate either of the two conservation laws we have stated. Some other restriction (or restrictions) must be operating.

Question 4.1.2 Make up some other illustrations of changes you are familiar with that go in only one direction and do not reverse themselves spontaneously even though the reversal would not violate either conservation of mass or momentum.

We shall find that all interactions leading to changes in state of objects or systems, involve “transfers” or “transformations” of an abstract quantity to which we give the name “energy.” (We use the word “quantity” because it turns out that we can assign numerical values.) Energy turns out to be a quantity that is conserved (in addition to mass and momentum), and we shall eventually be speaking of the law of conservation of energy. It is this conservation law that lies behind the impossibility of perpetual motion machines of the kind we invented earlier. Energy transformations are subject, however, to still another restriction in addition to conservation, and this other restriction (called the second law of thermodynamics) governs the *direction* of spontaneous change and indicates the directions that cannot occur spontaneously. Richard Feynman gives an especially lucid description of what “energy” means and does not mean.²

There is a fact, or if you wish, a law governing all natural phenomena that are known to date. There is no known exception to this law—it is exact so far as we know. The law is called the conservation of energy. It states that there is a certain quantity, which we call “energy,” that does not change in the manifold changes that nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says there is a numerical quantity which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a strange fact that when we can calculate some number and when we finish watching nature go through her tricks and calculate the number again, it is the same. . . .

. . . when we are calculating the energy, sometimes some of it leaves the system and goes away, or sometimes some comes in. In order to verify the conservation of energy, we must be careful that we have not put any in or taken any out. Second, the energy has a large number of different forms, and there is a formula for each one. These are: gravitational energy, kinetic energy, heat energy,

²Richard Feynman was one of the most eminent theoretical physicists of this century. He shared the Nobel Prize for development, in the late 1940s, of the theory of quantum electrodynamics, and he made many other significant contributions during his very fruitful career. At one time he taught the freshman physics course at California Institute of Technology through a series of lectures published as *The Feynman Lectures on Physics* by R. P. Feynman, R. B. Leighton, and M. Sands, Addison-Wesley Publishing Co., Reading, MA. 1963. The quoted passage is taken from vol. 1. The *Lectures* are wonderful reading for anyone interested in physics. They present an overview and reveal beauty and interrelationships in a manner very rarely found in science textbooks.

elastic energy, electrical energy, chemical energy, radiant energy, nuclear energy, mass-energy. If we total up the formulas for each of these contributions, it will not change except for energy going in and out.

It is important to realize that in physics today, we have no knowledge of what energy "is." We do not have a picture that energy comes in little blobs of a definite amount. It is not that way. However, there are formulas for calculating some numerical quantity, and when we add it all together it [always] gives . . . the same number. It is an abstract thing in that it does not tell us the mechanism or the reasons for the various formulas.

What we shall now set out to do is to *discover* what some of these various formulas are by re-examining some of the simple phenomena we dealt with earlier.

4.2 REVIEW OF IMPULSE, MOMENTUM, AND COLLISIONS

Let us review briefly some of the things we did and ideas we learned in studying momentum and its conservation because they suggest a fruitful parallel track for our present enterprise. In dealing with the rectilinear collision of two bodies A and B with masses m_A and m_B , with vector velocities before collision denoted by \vec{v}_{A1} and \vec{v}_{B1} , and vector velocities after collision denoted by \vec{v}_{A2} and \vec{v}_{B2} respectively, we learned that the connection between initial and final conditions was

$$m_A \vec{v}_{A2} + m_B \vec{v}_{B2} = m_A \vec{v}_{A1} + m_B \vec{v}_{B1} \quad (4.2.1)$$

i.e., that the total final vector momentum of the system is equal to the total initial vector momentum. Writing Eq. 4.2.1 in a slightly different way we have

$$m_A \vec{v}_{A2} - m_A \vec{v}_{A1} = -(m_B \vec{v}_{B2} - m_B \vec{v}_{B1}) \quad (4.2.2a)$$

or, the same statement in terms of changes (delta quantities),

$$\Delta(m_A \vec{v}_A) = -\Delta(m_B \vec{v}_B) \quad (4.2.2b)$$

which alters the previous verbal statement about total initial and final momenta to the equivalent statement that the change of momentum of one body is always equal and opposite to the change of momentum of the other.

Our other investigation started with Newton's second law

$$\vec{F}_{\text{net}} = m\vec{a} \quad (4.2.3)$$

and consisted of integrating both sides of Eq. 4.2.3 with respect to clock reading t :

$$\int_{t_1}^{t_2} \vec{F}_{\text{net}}(t) dt = m\vec{v}_2 - m\vec{v}_1 = \Delta(m\vec{v}) \quad (4.2.4)$$

For the special case of a *constant* net force, Eq. 4.2.4 reduces to

$$\vec{F}_{\text{net}} \Delta t = \Delta(m\vec{v}) \quad (4.2.5)$$

Question 4.2.1 Interpret both sides of Eqs. 4.2.4 and 4.2.5 as areas under \vec{F} versus t and \vec{a} versus t graphs, respectively, by sketching the corresponding graphs. For example, the left hand side of Eq. 4.2.5 is a rectangle with constant height \vec{F}_{net} along the F axis and with a base of length $t_2 - t_1$ along the t axis. What is the graph for the right hand side? The graphs for the two sides of Eq. 4.2.4 are similar, except that the force varies with clock reading in some arbitrary way of your own choice, and the areas are therefore not rectangular.

We gave the name “impulse delivered to the body by the force \vec{F}_{net} ” to the integral on the left hand side of Eq. 4.2.4 and the name “change of momentum of the body” to the quantity on the right hand side. We saw that Eqs. 4.2.2 and 4.2.4 were connected by Newton’s third law since, in the closed system of the two-body collision, the bodies exert equal and opposite forces on each other and therefore impart equal and opposite impulses and equal and opposite changes in momentum.

We saw in Chapter 1 that many of Newton’s predecessors and contemporaries were aware of conservation of momentum in collisions as an empirical fact without having Newton’s laws of motion at their disposal. (As a matter of fact, the empirical regularities observed in collisions were referred to as the “laws of motion” before the publication of Newton’s *Principia*.) Huygens, however, had gone a step further than other observers and had noted that, in *perfectly elastic* collisions, the quantity mv^2 was conserved, i.e., in a two-body collision of masses m_A and m_B with initial and final velocities v_1 and v_2 , respectively

$$m_A v_{A2}^2 + m_B v_{B2}^2 = m_A v_{A1}^2 + m_B v_{B1}^2 \quad (4.2.6)$$

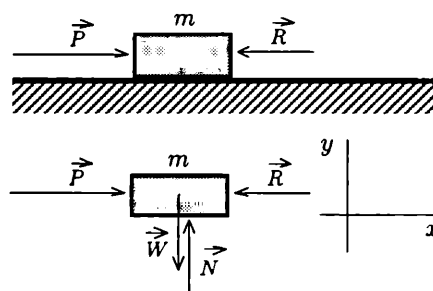
Huygens called the quantity mv^2 *vis viva* or “living force.” Huygens’s observation was purely empirical, that is, he found the numerical relation through measurement and not through any theory or derivation from underlying principles. We shall show in due course that there is a deep connection to underlying principles. We shall see that at least some *vis viva* “decays” or is “lost” in all real collisions while vector momentum is conserved in all collisions, whether elastic or inelastic.

Our experience with impulse and momentum and the connection to collisions suggests that it might be fruitful to examine some other integral of Newton's second law [similar to Eq. 4.2.4] in order to see whether there might be a connection to vis viva instead of to momentum.

4.3 HORIZONTAL ACCELERATION OF A SINGLE PARTICLE UNDER CONSTANT FORCES

The simplest case we can start with is that of a single particle confined to horizontal motion without friction. Let us imagine dealing with a frictionless puck of mass m on an air table or a glider of mass m on an air track as sketched in Fig. 4.3.1. We treat the object as a *particle*, meaning that we treat it as though all of it were concentrated at its center of mass since, for the time being, we are not concerned with its deformation, or its extension in space, or the molecules that comprise it.

Figure 4.3.1 Frictionless puck or glider of mass m subject to horizontal forces \vec{P} and \vec{R} .



We *define* our system to be the particle alone, and we take the forces \vec{P} and \vec{R} to be interactions with other systems not presently under consideration. For example, we ourselves might be exerting the push \vec{P} while \vec{R} might be exerted by some other person or object. The positive x -direction is shown by our choice of coordinate system. We take the instantaneous velocity \vec{v} to be positively directed. What can we say about changes in motion for various *displacements* $\Delta x = x_2 - x_1$? (Recall that, in dealing with momentum, we asked about changes in motion over *time* intervals $\Delta t = t_2 - t_1$.)

If the applied forces \vec{P} and \vec{R} are equal in magnitude, the velocity of the particle does not change, and we shall say that its state remains unaltered since we shall, in this context, not regard position as a significant state variable. (This will not always be the case. We shall see that when we displace an object vertically under the influence of gravity, or when we compress a spring, position *does* become a significant state variable.) The state *is* altered, however, if \vec{P} is not equal to \vec{R} . Let us suppose that the magnitude of \vec{P} is greater than that of \vec{R} so that the particle is accelerated in the positive x -direction, and let us first consider the simplest possible situation, namely that in which \vec{P} and \vec{R} , and therefore the acceleration \vec{a} , are all constant.

Since we shall be dealing with motion in one dimension only, let us simplify our notation by using the simple symbols P and R to represent the *magnitudes* of the vectors \vec{P} and \vec{R} , respectively. The net force $F_{x\text{ net}}$ in the x -direction and the acceleration a_x will carry plus or minus signs depending on whether they are directed to the right or to the left.

In Fig. 4.3.1 $F_{x\text{ net}} = P - R$, and we have, from Newton's second law:

$$P - R = ma_x \quad (4.3.1)$$

Since we are interested in what we can say about changes in motion for various displacements Δx , let us go back to the kinematic relation for constant rectilinear acceleration that connects change of velocity with Δx , namely

$$v_2^2 - v_1^2 = 2a(x_2 - x_1) = 2a\Delta x \quad (4.3.2)$$

where we have dropped the x subscripts on velocities and acceleration in order to avoid unnecessary clutter. Let us remember we are talking about the x -direction throughout this particular problem.

If we solve for a in Eq. 4.3.2 and substitute this expression for a_x in Eq. 4.3.1, we obtain

$$P - R = \frac{mv_2^2 - mv_1^2}{2\Delta x} \quad (4.3.3)$$

and, multiplying both side of Eq. 4.3.3 by Δx , we obtain

$$(P - R)\Delta x = \frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2 = \Delta\left(\frac{1}{2}mv^2\right) \quad (4.3.4)$$

Question 4.3.1 Interpret both sides of Eq. 4.3.4 as areas on graphs by sketching possible graphs just as you did for the impulse-momentum relation in question 4.2.1, except that the independent variable (along the abscissa) is now position x instead of clock reading t .

Note that the last step, that of multiplying both sides of Eq. 4.3.3 by Δx is not a matter of trickery or caprice. Look at what has been achieved in Eq. 4.3.4: On the right hand side we now only have terms that are intrinsic *properties* (i.e., state variables) of the system, namely the mass and velocity of the body. On the left hand side we have only the numbers that have to do with the action that external systems have *imposed* on the body, namely the forces and their associated displacement. Such a separation of characteristic quantities was not yet effected in the form of Eq. 4.3.3. Let us now make a very careful examination of Eq. 4.3.4, translating into words everything that it implies. (This is a process that is exceedingly important in *all* applications of physics, including engineering work and physics research, and you should practice making such interpretations at every opportunity on your own. This time guidance is being provided. You should follow every step of the following sequence with your own pencil and paper.)

Step 1 First examine and interpret the meaning of the algebraic sign on the right hand side of Eq. 4.3.4, i.e., the sign associated with the term $\Delta[(1/2)mv^2]$. Does this minus sign now have the same meaning it had when associated with direction of position changes and directions of velocities and accelerations along the x -axis? Try to answer this question for yourself before continuing to the following. [The instantaneous quantity $(1/2)mv^2$ is intrinsically positive because v^2 will be positive regardless of the algebraic sign of the velocity v . Therefore $(1/2)mv^2$ does *not* have *vector* properties. But the quantity $\Delta[(1/2)mv^2]$ *does* have algebraic signs. These signs, however, have nothing to do with the direction of motion. The algebraic sign is positive if the final velocity has a larger magnitude than the initial, and negative if it has a smaller magnitude, regardless of the direction of motion. Thus the algebraic sign of $\Delta[(1/2)mv^2]$ only tells us whether the property $(1/2)mv^2$ has increased or decreased in size during the applied action and no longer has anything to do with direction in space. We describe the difference that has arisen by saying that $(1/2)mv^2$ and $\Delta[(1/2)mv^2]$, like mass m , are both *scalar* rather than *vector* quantities.]

Step 2 On the left hand side of Eq. 4.3.4, we have the product of two quantities both of which have algebraic signs connected with positive and negative directions along the x -axis. $P - R$ is positive or negative depending on which force is bigger, i.e., on whether the net force is directed in the positive or negative x -direction. Δx is positive or negative depending on whether the particle was displaced in the positive or negative x -direction. How does the algebraic sign of this product of two vectors connect physically with the change in the magnitude of the scalar quantity on the right hand side? [(1) Suppose the net force $P - R$ and the displacement Δx are both positive. Which way is the particle going? What must be happening to the magnitude of its velocity? What must be happening to the *magnitude* of its $(1/2)mv^2$? Is the algebraic sign on the right hand side consistent with the sign on the left hand side? (2) Now suppose the net force $P - R$ is positive, meaning that P must be larger than R , but suppose that Δx is simultaneously negative. What must be happening, i.e., which way is the particle moving and what must be happening to the magnitude of its velocity in the light of the direction of the net force? What are the algebraic signs on both sides of the equation? (3) Examine and interpret, in a similar sequence, the signs for the situation in which $P - R$ is negative while Δx is positive. (4) Do the same for the case in which $P - R$ and Δx are both negative.]

Step 3 Consider the very simple case in which $R = 0$, and only the force P acts on the particle. Eq. 4.3.4 then becomes

$$P\Delta x = \Delta\left(\frac{1}{2}mv^2\right) \quad (4.3.5)$$

Interpretation: If there is no force opposing P , and if P and Δx are in the same direction, $P\Delta x$ is positive; the particle's $(1/2)mv^2$ is increased during

the displacement and the magnitude of this increase is equal to $P\Delta x$. This statement is by no means as trivial as it may sound initially. It says that we have discovered two entirely different ways of calculating the same number: If we know P and Δx , we can predict the change in the particle's $(1/2)mv^2$. If we know, from direct observation, the mass and initial and final velocities of the particle, we can calculate what constant force, acting over any given displacement, would have imparted the corresponding change in $(1/2)mv^2$. Furthermore, and most significantly, Eq. 4.3.5 tells us that, although a given external action $P\Delta x$ will impart very different velocities to bodies with different masses, it will always impart exactly the same *change* in $(1/2)mv^2$ to *any* body regardless of its mass!

Step 4 Let us rewrite Eq. 4.3.4 in a slightly different way:

$$P\Delta x = R\Delta x + \Delta\left(\frac{1}{2}mv^2\right) \quad (4.3.6)$$

Many ways of talking about the numbers in such relationships could conceivably be developed, but one particular language has evolved and taken root because of the way in which these numbers seem to be “preserved.” If, as we have already done in the preceding section, we think of $P\Delta x$ as an action that has been imparted to the body by the system exerting the force P (ourselves, for example), we note that this number has not disappeared. In Eq. 4.3.5, for example (or in Eq. 4.3.6 with $R = 0$), this number appears in the system under consideration (the particle) as the change in the intrinsic property $(1/2)mv^2$. We say that the action associated with $P\Delta x$ has been “transformed” into the change in $(1/2)mv^2$ or has been “conserved” as such a change.

If R is not equal to zero, we translate Eq. 4.3.6 with the statement “part of our action $P\Delta x$ is transformed into a change in $(1/2)mv^2$ of the system (the particle) to which the action was imparted,” and the remainder (equal in magnitude to $R\Delta x$) is “used up” or “transferred” in the sense that it becomes an action on *another* system, namely the one imparting the opposing force R .

To illustrate this idea more specifically, suppose the system exerting the force R is a nail which is displaced a distance Δx into a wall as the puck collides with the nail, imagined to be at the right hand side of Fig. 4.3.1. Consider the simplest case in which $P = 0$, and the puck possesses the quantity $(1/2)mv^2$ as it makes contact with the nail. R and Δx are oppositely directed so $R\Delta x$ is negative, and the magnitude of $(1/2)mv^2$ must decrease. (The puck comes to a stop when its $(1/2)mv^2$ reaches zero.) In this case, the fact that the object possessed an initial $(1/2)mv^2$ made it possible for it to impart an action $R\Delta x$, equal in magnitude to the initial $(1/2)mv^2$, to another system, namely the nail and the wall. We think of the numerical value of the initial $(1/2)mv^2$ as having been preserved or transformed in the interaction that has taken place.

Thus we begin to talk about these numbers as though an entity of some sort were involved because our whole structure of language, visualization, and

description seems to impel us to do so. Nevertheless we must discipline ourselves to remember that we are dealing with a way of talking about numbers obtained from certain formulas and not about an entity or a material object. (See the last paragraph of the quotation from Feynman in Sect. 4.1.)

Step 5 We must note what happens (or does not happen) when either P or Δx is zero. If the force is zero, the situation is trivial: The action associated with $P\Delta x$ is obviously zero, and the change in $(1/2)mv^2$ is also zero. Suppose, however, that Δx is zero while P is not zero. In this case we are exerting a force (and we can get very tired doing so), but regardless how big the force is, there is no “action” being imparted to change the value of $(1/2)mv^2$, and nothing happens in this respect, i.e., both sides of Eq. 4.3.5 remain zero.

Question 4.3.2 Interpret what Eq. 4.3.6 has to say about the case in which the puck moves with velocity v but P and R have the same magnitude. Is this consistent with Newton’s first and second laws?

We can see from the preceding discussion and interpretations that we are beginning to discover some of the formulas we set out to look for at the end of Sect. 4.1. The formulas we have encountered so far are ones that involve a force times its corresponding displacement as in $P\Delta x$ or $R\Delta x$, and the mass and velocity as in $(1/2)mv^2$. Let us now consider a slightly more general case than the one we have analyzed above: Suppose the force P acts at an angle θ to the horizontal instead acting horizontally (Fig. 4.3.2).

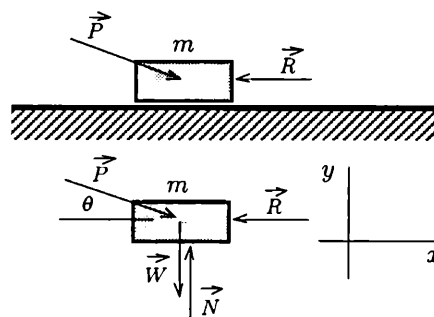


Figure 4.3.2 Frictionless puck of mass m subject to forces \vec{P} and \vec{R} , with force \vec{P} acting at angle θ to the horizontal.

In this case the net force accelerating the puck in the horizontal direction is $P \cos \theta - R$. Substituting into Newton’s second law, we have

$$P \cos \theta - R = ma_x \quad (4.3.7)$$

and, following exactly the same sequence we used earlier to obtain Eq. 4.3.4, we now obtain:

$$(P \cos \theta) \Delta x = R \Delta x + \Delta \left(\frac{1}{2} m v^2 \right). \quad (4.3.8)$$

Question 4.3.3 Fill in the algebraic steps between Eqs. 4.3.7 and 4.3.8.

Equation 4.3.8 indicates that the appropriate formula on the left hand side for the quantity that is “transformed” or “preserved” is not simply $P\Delta x$ (the entire force times the associated displacement as in Eq. 4.3.6) but $P \cos \theta \Delta x$, containing the *component* of the force in the *direction* of the displacement. Note, in particular, what Eq. 4.3.8 says about the calculation when the force P is *perpendicular* to the direction of displacement: The angle θ is then 90° , and the effect imparted by the force P is zero since $\cos 90^\circ = 0$. We can push very hard downward in the vertical direction, and get very tired doing so, but we produce no change in $(1/2)mv^2$.

Question 4.3.4 Consider the case in which $P = 0$, and the equations we have been examining reduce to

$$-R\Delta x = \Delta\left(\frac{1}{2}mv^2\right) \quad (4.3.9)$$

- (a) A body moves with velocity v_1 in the positive direction at the initial clock reading t_1 and is acted on by a passive opposing force R (such as friction, for example). How far will the body slide before coming to a stop? Don't just jump to a conclusion; solve the problem algebraically, step by step, paying *very* careful attention to the algebraic signs. (What, for example, is the value of v_2 ?) [Answer $\Delta x = +(1/2)mv_1^2/R$.]
- (b) Suppose that, with the same initial conditions as in (a), R , instead of being a passive force, is an active force exerted by your hand. The motion would then not cease at the instant the velocity v became zero but would be reversed. Describe, in words, the motion that would occur, and then show that Eq. 4.3.9 says the same thing in algebraic terms.

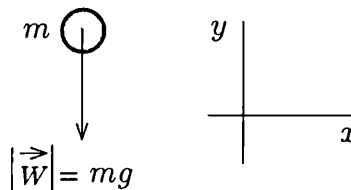
4.4 VERTICAL DISPLACEMENT OF A PARTICLE UNDER CONSTANT FORCES

Now let us see whether a similar set of relations involving “preserved” quantities exists in the case of vertical motion under the influence of gravity. We consider the idealized case, neglecting air resistance. First we look at a very simple relation that emerges directly from one of our familiar kinematic equations for uniformly accelerated motion, namely the one we already used as Eq. 4.3.2, except that now the magnitude of the acceleration is that of free fall, denoted by g , and the displacement of the particle is in the y -direction. We must therefore be very careful about the algebraic signs. The particle, the force acting on it, and the chosen coordinate system are shown in Fig. 4.4.1.

If we take g to denote the *magnitude* of the acceleration due to gravity and take the positive direction upward, the vector acceleration is negative, and our kinematic equation for free fall becomes:

$$v_2^2 - v_1^2 = 2g(y_2 - y_1) = 2g\Delta y \quad (4.4.1)$$

Figure 4.4.1 Particle of mass m free to accelerate under influence of its weight mg . Positive direction of y coordinate is upward.



Let us take the particle as the system under consideration. We multiply both sides of Eq. 4.4.1 by the mass m of the falling body and rearrange the equation into the pattern we established in Sect. 4.3, in which we had force times corresponding displacement on one side and intrinsic properties of the system on the other. (Remember that the *magnitude* of the force acting on the system in this instance is mg , the weight of the particle, and that its direction is negative.) We obtain:

$$-mg\Delta y = \frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2 = \Delta\left(\frac{1}{2}mv^2\right) \quad (4.4.2)$$

where the velocities denoted by v , are, of course, all in the y -direction.

Question 4.4.1 Interpret Eq. 4.4.2 very carefully, with emphasis on the meaning of the algebraic signs.

- (a) Note, for example, what the equation says about letting the particle go from rest at a height y_1 : Since the particle *falls*, Δy is *negative*, and the left hand side of the equation becomes *positive*; therefore the right hand side is also positive, which means that the numerical value of its $(1/2)mv^2$ increases as the system falls.
- (b) Suppose we start the particle in the upward direction with an initial upward velocity v_1 at height y_1 . What do you know actually happens to the value of $(1/2)mv^2$ on the way up? Does the equation agree with your physical expectation? Solve algebraically for the value of Δy_{\max} , the distance the particle rises between the instant of being projected at velocity v_1 and the instant it reaches the top of its flight. (What value of v defines the “top of the flight”?) Have you seen this relation previously? In what situations?
- (c) Show that Eq. 4.4.2 tells us the following: If a steel ball bearing is dropped from a height Δy above a level steel block, and *if* it experiences a *perfectly elastic* collision with the block, bouncing back with the magnitude of its vertical velocity v unchanged, the process of bouncing up and down and “converting” $mg\Delta y$ to $(1/2)mv^2$ and back again could go on indefinitely, without diminution of the maximum values of either of these numbers.

Note how certain numbers are preserved in this case, and note the essential similarity to the examination we conducted in Sect. 4.3: The left hand side of Eq. 4.4.2, $mg\Delta y$, contains the force acting on the particle (system) multiplied by the associated displacement. The magnitude of this quantity is “preserved.” If the force and displacement are in the *same* direction, the $(1/2)mv^2$ value

increases by exactly that amount; if the force and displacement are in *opposite* directions, the $(1/2)mv^2$ decreases by exactly that amount. Thus the $(1/2)mv^2$ value does not really “disappear” on the way up; it is “stored” as $mg\Delta y$ by virtue of the gravitational interaction between the particle and the earth. This kind of preservation of numbers was well known to natural philosophers as early as the 17th century, but its broad extension to a wider range of natural phenomena did not occur until the middle of the 19th century, largely because it took such a long time to recognize the very subtle connection to thermal interactions and transfer of heat.

Question 4.4.2 Proceed, on your own, to extend the discussion to the case in which we add a constant external vertical force P acting upward on the particle (system) in Fig. 4.4.1. Imagine the force P to be exerted by your own hand, and take Δy to represent the displacement that takes place during the interval force P is exerted on the particle.

- (a) First re-draw the diagram, adding this force. Then show that one obtains the relation:

$$P\Delta y = mg\Delta y + \Delta\left(\frac{1}{2}mv^2\right) \quad (4.4.3)$$

- (b) Interpret what Eq. 4.4.3 has to say about the preservation of the various numbers for cases such as (1) $P > mg$; (2) $P < mg$; (3) $P = 0$; and (4) $P = mg$. In each case, examine what happens if the initial velocity v_1 of the particle is taken to be zero, or upward, or downward. In part (4) consider the case in which the particle has a very, very small upward or downward velocity which remains constant during the displacement. Be sure that your interpretations of the equation are consistent with what you know actually happens physically in these various instances.
- (c) Interpret each one of the terms in Eq. 4.4.3 as an area under a force versus position y diagram.

4.5 DISPLACEMENT OF A PARTICLE UNDER ACTION OF A VARYING FORCE

A discerning reader might offer the objection that the analyses carried out in the preceding two sections have really told us nothing new that we did not already know from Newton’s second law and the very special and restricted case of constant force and uniformly accelerated motion (Eqs. 4.3.2 and 4.4.1). He or she might point out, quite legitimately, that we simply rearranged an algebraic combination of these equations and developed a special jargon about the results. And the objection would be well taken if that were as far as the investigation could be carried. The analyses in Sects. 4.3 and 4.4 are, however, illustrative of a deep and general relationship that does not depend on the assumption of a constant force and uniform acceleration. The basic ideas turn

out to be valid, and numbers are preserved in exactly the same way, whether or not the forces are constant.

The course of the more general analysis is hinted at by the relationship we have seen among the *areas* on graphs in our earlier study questions. In calculus we have established a connection between integrals and areas under graphs of different shapes, graphs in which the ordinate is not a constant. We now proceed to exploit this connection. Suppose we know how the force acting on a particle (the particle being taken as our system) varies with position of the particle, i.e., we have a graph or an equation for the force as a function of position. Let us consider the case in which the particle is being accelerated in the x -direction by a varying net force denoted by $F_{x\text{ net}}(x)$, the notation indicating explicitly that the force is now being taken as a function of position x . A possible force versus position graph is shown in Fig. 4.5.1.

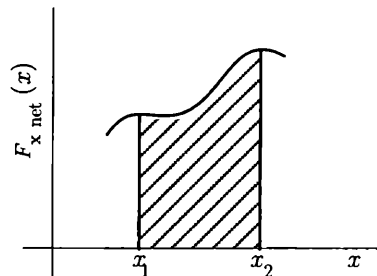


Figure 4.5.1 Graph of varying net force, $F_{x\text{ net}}(x)$, versus position x .

The shaded area under the graph in Fig. 4.5.1 is given by

$$\text{Area} = \int_{x_1}^{x_2} F_{x\text{ net}}(x) dx \quad (4.5.1)$$

and, if the force is constant, the area becomes a rectangle and is given by

$$\text{Area} = F_{x\text{ net}}(x_2 - x_1) = F_{x\text{ net}} \Delta x \quad (4.5.2)$$

Suppose our system is a particle of mass m being accelerated in the x -direction by the net force $F_{x\text{ net}}(x)$. By Newton's second law

$$F_{x\text{ net}}(x) = ma_x \quad (4.5.3)$$

If we integrate both sides of Eq. 4.5.3 with respect to position x , we are equating the two areas represented by each side:

$$\int_{x_1}^{x_2} F_{x\text{ net}}(x) dx = \int_{x_1}^{x_2} ma_x dx \quad (4.5.4)$$

We cannot evaluate the area on the left hand side unless we know the entire history of the variation of the force, instantaneous position by instantaneous position as in Fig. 4.5.1. If we have the equation of the graph, we might be able to integrate the equation. If we have the graph itself without an equation,

we can find the area numerically by counting unit squares. But we must have the graph one way or another. Let us examine what we can do, however, with the right hand side by playing with the meaning of the acceleration a in terms of derivatives. The object is to try to find a form for which we might know the antiderivative, i.e., a form that we can integrate directly.

Returning to the original definition of acceleration, we have

$$a \equiv \frac{dv}{dt} \quad (4.5.5)$$

and, by the chain rule for differentiation

$$\frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} \quad (4.5.6)$$

Since the very last term on the right hand side of Eq. 4.5.6 is simply the velocity v , combining Eqs. 4.5.5 and 4.5.6 gives

$$a = v \frac{dv}{dx} \quad (4.5.7)$$

a new expression for acceleration, and we know what has to be differentiated to give $v(dv/dx)$. It is $v^2/2$, since, again applying the chain rule

$$\frac{d}{dx} \left(\frac{v^2}{2} \right) = \frac{d}{dv} \left(\frac{v^2}{2} \right) \frac{dv}{dx} = v \frac{dv}{dx} \quad (4.5.8)$$

Putting the antiderivative $v_x^2/2$ into the right hand side of Eq. 4.5.4, we have

$$\int_{x_1}^{x_2} F_{x \text{ net}}(x) dx = m \left[\frac{v_x^2}{2} \right]_{v_{x1}}^{v_{x2}} = \frac{1}{2} m v_{x2}^2 - \frac{1}{2} m v_{x1}^2 = \Delta \left(\frac{1}{2} m v_x^2 \right) \quad (4.5.9)$$

If $F_{x \text{ net}}(x)$ happens to result from the combination of two forces such as P and R in Fig. 4.3.2, and if both of these forces are functions of x , Eq. 4.5.9 becomes

$$\int_{x_1}^{x_2} P(x) \cos \theta dx = \int_{x_1}^{x_2} R(x) dx + \Delta \left(\frac{1}{2} m v_x^2 \right) \quad (4.5.10)$$

These are very powerful results. We are no longer confined to constant forces and can deal with the numbers that are preserved in *any* circumstance. The basic result is that the change imparted to the $(1/2)mv^2$ value for the particle is always equal to the area under the force versus position curve for the given displacement regardless of how the force varies!

Question 4.5.1 Interpret Eq. 4.5.10 by considering some cases of horizontal motion just you did in Sect. 4.3. The interpretations are virtually the same as those in Sect. 4.3, but the result is now much more general because we are no longer confined to constant forces. Show as part of your interpretation that, even

though we may know nothing of how the force varies over a given displacement, if we know the magnitude of the change in $(1/2)mv^2$, we at least know the area under the corresponding force versus position graph, and, if we know the value of the displacement, we can calculate the average value of the force that must have acted on the particle. (You can assist your grasp of the idea involved by going back to Sect. 1.10 where you did exactly the same thing with impulse and momentum.)

Question 4.5.2 Suppose you are carrying a suitcase along level ground at uniform velocity. You are exerting a vertically upward force of magnitude $P = mg$ on the suitcase. Let y and x represent the vertical and horizontal coordinates, respectively, and let P_y and P_x represent the vertical and horizontal components of P , respectively. What are the numerical values of $P_y\Delta y$ and $P_x\Delta x$ for this situation? What effect would this force have on the value of $(1/2)mv^2$ of the particle? Be sure to explain your reasoning.

4.6 DISPLACEMENT OF A PARTICLE AGAINST THE RESTORING FORCE OF A SPRING

Consider the situation illustrated in Fig. 4.6.1: We exert a force P toward the right on a frictionless puck of mass m and compress a spring of negligible mass located between the puck and the wall. The spring is assumed to obey Hooke's law, i.e., if we take the origin of coordinates at the end of the spring when the spring is relaxed, the magnitude of the force F_{xs} exerted by the spring on the puck when the end of the spring is displaced to position x is equal to kx , where k is called the "spring constant."³ We assume the magnitude of P to be some arbitrary function of x , i.e., $P(x)$.

We take the puck (particle) as the system to be considered. Then, from Fig. 4.6.1, we have

$$F_{x\text{ net}}(x) = P(x) - F_{xs} = P(x) - kx \quad (4.6.1)$$

where $F_{x\text{ net}}(x)$ carries algebraic signs and $P(x)$ and kx are magnitudes.

If we displace the particle from $x = 0$, the relaxed position of the spring, to some final position x , integrating both sides of Eq. 4.6.1, gives

$$\int_0^x F_{x\text{ net}}(x)dx = \int_0^x P(x)dx - \int_0^x kx dx = \int_0^x P(x)dx - \frac{1}{2}kx^2 \quad (4.6.2)$$

³We are considering the case in which the mass of the spring is very, very small (negligible) compared to the mass of the block. This means that the forces necessary to accelerate the coils of the spring in their displacements are very, very small relative to F_{xs} and that the $(1/2)mv^2$ values for the spring coils are also negligible. If we were to take the mass of the spring and the motion of its parts into account, we would be dealing with a far more complicated problem than we are now prepared to tackle. In particular, the force F_{xs} would be larger than kx by amounts difficult to establish.

or if, slightly more generally, we integrate from position x_1 to position x_2

$$\int_{x_1}^{x_2} F_{x \text{ net}}(x) dx = \int_{x_1}^{x_2} P(x) dx - \left[\frac{1}{2} kx_2^2 - \frac{1}{2} kx_1^2 \right] \quad (4.6.3)$$

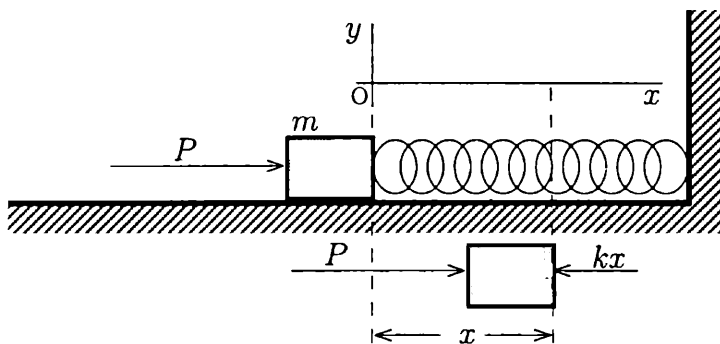


Figure 4.6.1 Horizontal force of magnitude P is exerted on frictionless puck of mass m . As puck is displaced to the right from initial relaxed position of end of spring at $x = 0$ to position $x = x$, the magnitude of the horizontal force exerted by the spring on the puck, acting toward the left, increases from 0 to kx . The mass of the spring is taken as vanishingly small relative to that of the puck. The force diagram for the puck has omitted the balanced vertical forces to avoid clutter.

If we put the results in Eqs. 4.6.2 and 4.6.3 into the general relation derived in Eq. 4.5.10, we obtain

$$\int_0^x P(x) dx = \frac{1}{2} kx^2 + \Delta \left(\frac{1}{2} mv_x^2 \right) \quad (4.6.4)$$

and

$$\int_{x_1}^{x_2} P(x) dx = \frac{1}{2} kx_2^2 - \frac{1}{2} kx_1^2 + \Delta \left(\frac{1}{2} mv_x^2 \right) \quad (4.6.5)$$

Question 4.6.1 Be sure to work through the substitution leading to Eqs. 4.6.4 and 4.6.5 so as to verify the algebraic signs.

If the particle is released from rest ($v_{x1} = 0$) at an initial position x_1 , (i.e. with the spring compressed by a displacement x_1 of the end from its relaxed position) and is allowed to move freely (i.e., with $P = 0$), Eq. 4.6.5 becomes:

$$0 = \frac{1}{2} kx_2^2 - \frac{1}{2} kx_1^2 + \left(\frac{1}{2} mv_{x2}^2 - 0 \right) \quad (4.6.6)$$

or

$$\frac{1}{2} kx_2^2 = \frac{1}{2} kx_1^2 - \frac{1}{2} mv_{x2}^2 \quad (4.6.7)$$

Question 4.6.2 Interpret Eqs. 4.6.5 and 4.6.7 in careful detail, following the pattern outlined in question 4.4.2. Show that very similar conservation relationships obtain. There are, however, important differences in the character of the motion that can take place: If we drop a body vertically, from rest, through a displacement Δy , it acquires a quantity of $(1/2)mv^2$ equal to the magnitude of $mg\Delta y$ and will continue dropping and acquiring still more $(1/2)mv^2$ if the fall is not interrupted. If in the case of the spring, however, we release the particle from a compressed spring position $+x_1$ (Fig. 4.6.1), it will be accelerated toward the left, and, on returning to $x = 0$, the particle will have acquired an amount of $(1/2)mv^2$ equal to $(1/2)kx_1^2$. It will then continue moving to the left, but the acceleration will be directed toward the right because of the stretching of the spring. In the case of free fall, there was no such reversal of the direction of acceleration.

Show that the equations tell us that the particle will continue moving to the position $x = -x_1$, at which point the direction of motion will be reversed. In the absence of friction, this oscillation would continue indefinitely (as in the case of the bouncing steel ball in question 4.4.1) with the maximum values of $(1/2)kx^2$ and $(1/2)mv^2$ being continually interchanged. The motion is “symmetrical” around the position $x = 0$. What mathematical characteristic of Eq. 4.6.7 accounts for this symmetry? Solve Eq. 4.6.7 for v_{x2} , and interpret the result very carefully in words. At what position or positions is v_{x2} zero? At what position or positions does it have its largest value?

In the preceding discussion, we took the particle (puck) *alone* as the system whose state was to be analyzed and described. The puck was subjected to changes in state by two interactions, the force P , exerted by us or some other active agent, and the force of magnitude kx exerted by the spring. Let us begin to explore what considerations arise when we elect to take a different system.

Suppose we now take the system under consideration to be the combination of the spring and the puck, instead of the puck alone, with force P still acting from the outside the system. Now we have only one external interaction influencing the state of the system instead of two, the other interaction being internal. Suppose we put the system into some initial state, such as position x_1 of the puck, and consider what follows if $P = 0$. The system is now isolated in the sense that none of the external forces acting on it (the weight of the puck, the normal force exerted by the table on the puck, the normal force exerted by the wall on the spring) undergo any displacements that change the state of the combination. (The force exerted by the wall on the spring is not the force that displaces the end of the spring and effects the compression. Compression is imposed by the force P .) In this system (puck and spring, with $P = 0$) all the interactions resulting in changes of state take place *internally*, or within the system, rather than between the system and surrounding systems. Perceiving such distinctions between different choices of system will play an important role in our subsequent analyses.

4.7 VOCABULARY: WORK AND KINETIC ENERGY

At this point it becomes appropriate to say something about terminology. We have deliberately refrained from naming the new concepts in order to focus attention on the ideas that emerge: the formulas, the numerical relationships, and the operational definitions. We have been operating under the precept “idea first and name afterwards.” Many of the names we select in physics are metaphors borrowed from everyday speech, such as “force,” “mass,” “work,” “energy,” “charge,” “current,” “potential,” and various others. Our modes of thought are intimately affected, and sometimes even determined, by the structure of our language and vocabulary. Borrowing metaphors from ordinary speech has more than once involved science in a morass of misunderstanding and confusion from which it was extricated only by an act of genius. Being careful to develop the idea before bringing forth a name helps to avoid such confusion.

Late 17th and early 18th century scientific literature contains polemics concerning the proper definition of “force.” Huygens and Newton held that force was properly measured in terms of the “change in quantity of motion” which, to them, meant changes in $m\vec{v}$. Leibniz, on the other hand, argued that force was more properly connected with changes in vis viva or “living force” (mv^2). Needless conflict arose over the use of the same name for what were clearly different operational concepts. Until the middle of the 19th century, considerable confusion prevailed because of the indiscriminate, interchangeable use of the terms “force” and “energy.” In our modern terminology, the issue is clarified by reserving the word “force” for the concept you have seen defined operationally in the development of Newton’s laws of motion. The name “energy” is introduced as a general term to describe the diverse numbers and formulas we have been discovering in this chapter. These numbers are interesting because they reveal an abstract quantity that seems to be preserved in the midst of complex changes.

We give the name “work” to numbers calculated from products of force and displacement such as $P\Delta x$ or integrals of varying forces, such as $\int_{x_1}^{x_2} P(x)dx$. We speak of the “work done by force P in the displacement from position x_1 to position x_2 ,” and we speak of “work” as a measure of energy *transferred* from one system to another through an interaction involving forces and the displacements these forces undergo.

In the more general situation (such as that considered in Eq. 4.3.8, where P might act at an angle θ to the displacement, we have seen that the quantity conserved is *not* $\int_{x_1}^{x_2} P(x)dx$, but is, rather, $\int_{x_1}^{x_2} P(x) \cos \theta dx$. Thus, in order to obtain numbers that are conserved, it is necessary to define work by the latter integral and not the former; that is, we must not use the magnitude of the force itself in the arithmetical calculation, but the *component* of the force in the direction of its displacement, otherwise we are not calculating a conserved quantity.

In the still more general terms of vector notation, this means we must calculate the vector dot product between the force and the corresponding displacement. We say that the appropriate definition of work W done by a force $P(s)$ in a displacement from some general position s_1 to position s_2 must be

$$W \equiv \int_{s_1}^{s_2} \vec{P} \cdot d\vec{s} \quad (4.7.1)$$

in order to yield numbers that are conserved. (Note that the dot product is also called the “scalar product” of vectors because it yields a scalar quantity. We have seen work to be a scalar quantity, and Eq. 4.7.1 is consistent with that insight.)

We have encountered situations in which forces must be said to do zero work because they do not produce any displacement in the system on which they act (e.g., the forces exerted on the suitcase when carrying the latter at uniform velocity, the weight and normal force exerted on the puck in Fig. 4.6.1, the force exerted by the wall on the spring in Fig. 4.6.1). Such forces are referred to as “zero work” forces, and we shall encounter many of them in future problems. For example, as a preview, in rather more subtle situations such as those in which (1) the ground exerts a force on us when we walk or run, (2) the ground exerts a horizontal frictional force on the accelerating car, (3) the wall exerts a force on us when, while standing on roller skates, we push ourselves away from the wall—in all these instances the forces exerted by the ground and the wall are zero work forces! They produce no energy changes that are conserved. As just remarked, this is a subtle matter; you would do well to start thinking about it now so as to be prepared for deeper consideration later on. Do not feel diminished if you find initial difficulty with this idea; you will be in good company.

We may push on an unyielding wall for a long time and become very tired from doing so, but, in the sense of our physical definition, we have done zero work. As a scientific concept, “work” is not necessarily a measure of our exhaustion—an illustration of the fact that metaphors do not transform literally to the new context and, unless clearly separated from the original context, can easily introduce elements of confusion into our thinking and reasoning.

Work integrals emerge with algebraic signs that have nothing to do with directions in space of the original force vectors. The sign is *positive* if force and displacement are in the *same* direction; *negative* if they are in *opposite* directions. In the former case (positive), we speak of the “work done *by* the system (which is exerting the force) on some other system,” or, more simply, as “work done by the force.” If the work quantity is negative, we speak of the system exerting the force as *receiving* work from the other system on which it acts, or, sometimes, as work done *against* the force in question.

The quantity $(1/2)mv^2$, which increases or decreases as work is done by or against a net force, is given the name “kinetic energy,” and the relation

obtained in Eq. 4.5.9:

$$\int_{x_1}^{x_2} F_{x \text{ net}}(x) dx = \frac{1}{2} m v_{x2}^2 - \frac{1}{2} m v_{x1}^2 = \Delta \left(\frac{1}{2} m v_x^2 \right) \quad (4.7.2)$$

is called the work-kinetic energy theorem. Translating this theorem into words, we say that “all the work done by a net force accelerating a body is transformed into a change in kinetic energy (KE) of the moving body.” Note that kinetic energy is a property of the moving object; it is a state variable in that it is one of the numbers that describes the state of the system.

“Energy” is *not* a substance, fluid, paint, or fuel which is smeared on bodies and rubbed off from one to another. As Feynman is quoted as saying at the start of this chapter, we have no idea what energy “is,” and “is” is the wrong word to use in such a connection. We use the term “energy” to denote an abstract construct, referring to numbers that are calculated in certain prescribed ways by means of formulas that we *discover*. These formulas are arrived at by combinations of theory and experiment, and are deemed important and useful because the resulting numbers are found to be preserved and to reveal remarkably simple regularities in seemingly very diverse and unrelated physical phenomena.

4.8 POTENTIAL ENERGY

In Sect. 4.4, we found that, in the case of vertical motion of an object under gravity in the absence of frictional forces, energy appears to be “stored” in the sense that the numbers represented by $mg\Delta y$ and $(1/2)mv^2$ are converted back and forth into each other as the object rises and falls.

In Sect. 4.6 we found that, in the case of motion of the puck under the influence of an elastic spring in the absence of frictional forces, energy appears to be “stored” in the sense that the numbers represented by $(1/2)kx^2$ and $(1/2)mv^2$ are converted back and forth into each other as the puck oscillates back and forth on its frictionless plane.

In the gravitational case, the interaction involves both the earth and the object whose elevation is changing. If a particle has initial velocity v_1 directed vertically upward, its KE $(1/2)mv^2$ decreases as it rises and has the instantaneous value of zero at the instant it reaches the top of its flight. But the initial energy is said to be “stored” because it does not simply disappear (as it seems to do if the body slides to a stop along a floor with friction); the KE is “recovered” when the body falls back to its initial position even though the velocity is then in the opposite direction. Alternatively, if the body is caught at the top of its flight, it can be rigged, through a string and pulley arrangement, to elevate another identical body to the position it reached while it is itself being returned to its starting position. Under the influence of air resistance (a form of friction), however, the entire amount of KE is not preserved or stored;

it seems to “leak away” and is permanently lost to the system. Such “lost” energy is said to have been “dissipated.”

In the case of the frictionless puck and the spring (Sect. 4.6), an initial KE imparted to the puck appears to be “stored” in compression of the spring, which “un-compresses” after the maximum displacement and restores the initial KE of the puck. As the oscillation described in Question 4.6.1 continues (in the absence of friction), the KE of the puck is alternately stored and restored as the spring is compressed and extended and maintains the fixed value initially imparted. When friction is present between the puck and the table, the oscillation eventually comes to a complete stop, and all the energy originally imparted is dissipated.⁴

Energy stored in the earth-body system as the body is elevated, energy stored in the spring as it is compressed or extended, is given the name “potential energy” (PE for short). This name is used only when the stored energy is recoverable in other forms as in the cases just cited. (The energy dissipated in frictional interactions is not described as potential energy.) Changes in PE are associated with integrals such as

$$\int_{y_1}^{y_2} mg dy = mg(y_2 - y_1) = mg\Delta y \quad (4.8.1)$$

and

$$\int_{x_1}^{x_2} kx dx = \frac{1}{2}kx_2^2 - \frac{1}{2}kx_1^2 = \Delta\left(\frac{1}{2}kx^2\right) \quad (4.8.2)$$

These two calculations suggest speaking of different “forms” of potential energy: gravitational in the first case and elastic or spring-like in the second. Subsequently we shall speak of electric, magnetic, chemical, nuclear, and other forms. Forces that permit storage of PE (as do gravitational, elastic, and electrostatic forces) are called “conservative” forces while friction is called a “non-conservative” or “dissipative” force.

Having adopted these names, it is convenient to say that the PE of a system *increases* [$\Delta(\text{PE})$ positive] when a weight is raised or a spring is compressed or extended and that the PE of a system *decreases* [$\Delta(\text{PE})$ negative] when a weight is lowered or a spring is allowed to relax. Examine the calculations: If we have to put work *into* the system to effect the change under consideration,

⁴The conservation relation we discern in the idealized situations, in which friction is imagined to be absent, is referred to as “conservation of mechanical energy.” The usefulness of this idea is extremely limited because of its highly restricted applicability. The great breakthrough on fully understanding the generality of energy conservation came with the realization in the 19th century that heat was not an imponderable substance (as some models held) but was another form of energy. The energy dissipated under the influence of friction does not really disappear. It changes the state of the rubbing (interacting) bodies by elevating their temperatures. When double the amount of work or KE is dissipated, the thermal effects are exactly doubled. We call the latter form “thermal energy.” We shall explore these ideas in more detail in a later section.

$\Delta(\text{PE})$ of the system is positive. If we must receive work *from* the system to effect the change, $\Delta(\text{PE})$ of the system is negative. We calculate these changes in PE by calculating the amount of work we would have to do (or receive) in producing the given displacement without appreciable change in the KE of the system.

One must be careful about localizing PE or thinking of it as a property of a single particle (the way KE is indeed a property of the single moving particle, once we have specified the frame of reference for measurement of velocity.) In the case of extended, deformable objects such as the compressed spring or a compressed gas in a cylinder, PE is localized in the object (the spring or the gas) and can be considered a property of the object, but such extended objects are not single particles and are really systems of interacting particles (atoms and molecules.) In the case of the body being elevated against the gravitational pull of the earth, PE is certainly not localized in the body, nor is it a property of the body alone. In this case the PE is a property of the body-earth *system*. Without the presence of *both* interacting bodies, there would be no PE manifest at all.

In cases of conservative forces, it turns out that $\Delta(\text{PE})$ between any two specified positions depends only on the location of the positions and is independent of the *path* traversed from one position to the other. Let us illustrate how this works out to be the case. Consider the situation in Fig. 4.8.1. We push a block up a frictionless inclined plane from position s_1 to position s_2 for a net elevation gain of $(y_2 - y_1)$.

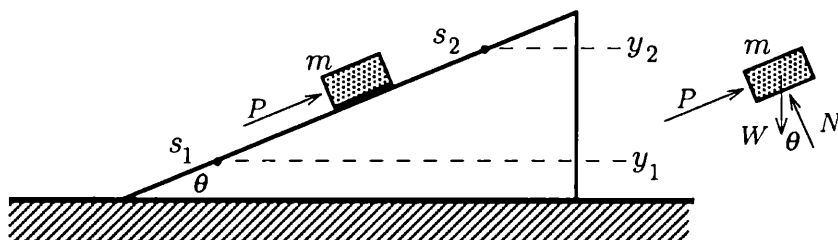


Figure 4.8.1 Block of mass m , treated as a particle, is pushed up the frictionless inclined plane from lower position s_1 to upper position s_2 with vanishingly small KE. The elevation change is from y_1 to y_2 . Position of the block is defined as position of the center of mass.

Let us take the positive direction for positions s to be up the plane. To push the block up the plane in the absence of friction and with negligible addition of kinetic energy, we must exert a force, parallel to the plane, of magnitude

$$P = mg \sin \theta \quad (4.8.3)$$

and, in effecting the displacement from s_1 to s_2 , we do the amount of work

$$W = \int_{s_1}^{s_2} (mg \sin \theta) ds = (mg \sin \theta)(s_2 - s_1) = (mg \sin \theta) \Delta s \quad (4.8.4)$$

Since the change in PE is equal to the work done on the system without addition of KE, $\Delta(\text{PE})$ of the block-earth system is given by

$$\Delta(\text{PE}) = (mg \sin \theta) \Delta s \quad (4.8.5)$$

The geometry of the figure shows a very simple connection between the elevation change Δy and the position change Δs along the plane:

$$\Delta y = (\Delta s) \sin \theta \quad (4.8.6)$$

and, using this relation to replace Δs in Eq. 4.8.5, we obtain

$$\Delta(\text{PE}) = mg \Delta y \quad (4.8.7)$$

Eq. 4.8.7 tells us that the change in PE, effected by pushing the block through the longer distance along the frictionless plane, is identical with the change we *would* have produced had we elevated the block vertically from y_1 to y_2 . If we let the block go from s_2 or y_2 , it would acquire the same KE (in the absence of friction) whether it slid down the plane to s_1 or fell directly vertically to y_1 .

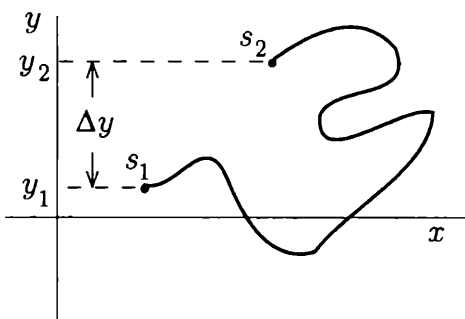
As far as pushing the block up the plane is concerned, the benefit of the gentle slope is that we exert a smaller force in order to displace the block than we would exert to displace it vertically, and we take advantage of this effect in using ramps for elevating heavy objects. To get to s_2 , however, we must push over a distance just so much greater than the direct vertical displacement that the work done is exactly the same by either route. If, for example, it took less work to push the block up the plane than it took to raise it vertically, we could set up a machine in which the block was pushed up the plane and then allowed to fall vertically from y_2 to y_1 . The KE of the block at y_1 would then be greater than the amount of work we put into the system, and we would have an inexhaustible source of energy for doing various useful things such as pumping water, raising other objects, generating electricity, and so forth. We begin to see that the conservation law we are gradually unraveling must be that which forbids, in a very general sense, such "getting something for nothing."

The result we obtained in the special, simple case of the inclined plane is susceptible to broad generalization: Suppose a particle is transported from position 1 (at some initial level y_1) to position 2 along some arbitrarily chosen, complicated curved path such as that shown in Fig. 4.8.2.

We can imagine cutting the curve up into a sequence of short inclined planes, each one as in Fig. 4.8.1, and evaluating $\Delta(\text{PE})$ as the limit of the

sum of quantities of work done along each little plane as the number of such subdivisions increases without limit, i.e., carrying out the integral along the path from position 1 to position 2. The increments will sometimes be positive, sometimes negative, and the final sum will still be $mg\Delta y$ regardless of the actual path followed.

Figure 4.8.2 Particle transported from position s_1 to position s_2 relative to the ground with vanishingly small KE along an arbitrarily complicated path. Net change in elevation is Δy .



This same idea can be extended to any conservative force (such as electric and magnetic forces). The change in PE between two points depends only on the locations of the points and is completely independent of the path followed in going from one point to the other. In contrast, suppose that points s_1 and s_2 in Fig. 4.8.2 were locations on a floor over which we push a box, subject to the opposing frictional force, without appreciable change in KE. In this case, the work we do depends on the path we follow between points s_1 and s_2 . We do the least amount of work along the shortest path and increasing amounts of work along longer paths. Furthermore, the work done does not seem to have been preserved; the box will not “fall” back from position s_2 to position s_1 if it is “released” at the latter position, nor will it acquire KE after it is released.

This distinction—path independence in some cases and path dependence in others—is eventually made the definition of “conservative forces” on the one hand (path independence for work done on the system between two positions) and “non-conservative (or dissipative) forces” on the other (path dependence for work done). We have just illustrated the fact that the force of gravity is a conservative force while friction is non-conservative. One must, however, be sensitive and very careful about the vocabulary. It is the *work done* by forces that is conserved or not conserved. Forces themselves are not conserved quantities as are mass, momentum, and energy. The terms “conserved” and “conservative” do not have the same meaning in this context.

All the equations we have derived are statements of relationship among energy *changes*. These energy changes are associated with alteration of the condition or state of the system being described, as defined by changes in motion and position of bodies. We have no way of defining reference levels for energy that are zero in some absolute sense; we can only designate some arbitrarily chosen state as corresponding to a zero level and calculate increases or decreases of energy relative to this reference state.

For example, we can designate the PE of a body as zero at any arbitrary elevation y_0 , not necessarily at ground level. Then at any elevation above y_0 , we would say the value of PE of the earth-body system would be positive, while at any elevation below y_0 , we would say the PE was negative. Negative values of PE would not mean that something was missing or owing; they would simply reflect the fact that we were talking about a position below the zero reference level. A body falling from such a location would be gaining KE while the PE of the system was becoming increasingly negative.

Consider a system subject only to conservative forces (such as gravity, or spring forces, or both simultaneously as in the case of an object hanging on a spring) with no frictional effects. Let us suppose the system to be isolated in the sense that no external forces (such as P in some of our earlier analyses) are doing any work on the system. For such a case, the results of the equations we derived in the preceding sections can be summarized in the statements of Eqs. 4.8.8a and 4.8.8b, which say the changes in PE and KE for the closed system must add to zero. In other words, an increase in PE is accompanied by an equal decrease in KE and vice versa.

$$(\text{PE})_2 - (\text{PE})_1 + (\text{KE})_2 - (\text{KE})_1 = 0 \quad (4.8.8a)$$

$$\Delta(\text{PE}) + \Delta(\text{KE}) = 0 \quad (4.8.8b)$$

Alternatively, Eq. 4.8.8a can be written in the form

$$(\text{PE})_2 + (\text{KE})_2 = (\text{PE})_1 + (\text{KE})_1 \quad (4.8.9)$$

where each side of Eq. 4.8.9 can be thought of as representing a “total” energy at conditions 1 and 2, respectively. If we take H to represent the initial total energy (the sum of the initial values of PE and KE), Eq. 4.8.9 can be expressed as

$$\text{PE} + \text{KE} = H \quad (4.8.10)$$

which says that, if we start the system with a total energy H , the sum of the subsequently changing values of PE and KE must remain equal to H , instant by instant. Eqs. 4.8.8a through 4.8.10 all say exactly the same thing in slightly different ways.

4.9 UNITS AND DIMENSIONS

Quantities of work and changes in potential energy are calculated from products of force and displacement, the dimensions of which we denote by F and L , respectively. Thus the dimensions of work are $F \times L$, and the units would be “newton meters” (N m) in the SI system and “foot pounds” (ft lb) in the British system. The name “joule” (J) is given to the combination newton-meter, and there is no special name for the British unit.

If we use the symbols M , L , and T to denote the dimensions of mass, length, and time, respectively, kinetic energy, defined by the formula $(1/2)mv^2$, has dimensions ML^2/T^2 . Work, PE, and KE are all quantities that, in accordance with our algebraic derivations, are to be combined with each other by addition and subtraction. For this operation to make sense, the dimensions of work and KE must be identical.

Question 4.9.1 Verify the preceding statement by showing first that the dimensions of force (since it is equal to mass times acceleration) must be ML/T^2 . Then complete the analysis to show that the dimensions of work and KE are indeed identical, meaning that both can be measured in joules (J).

4.10 PERFECTLY INELASTIC COLLISIONS

To illustrate the significance of some of the ideas we have been developing, let us return to the perfectly inelastic rectilinear collision of two bodies that we began discussing in Sects. 1.3 and 1.5 (the two bodies stick together on impact and go off with a common velocity). Let us take body B, with mass m_B , to be initially stationary (i.e., $v_{B1} = 0$), and body A, with mass m_A , having velocity v_{A1} at the instant of collision. The two bodies stick together and go off with the velocity v_2 . Conservation of momentum requires that the momentum of the system after collision must equal the momentum before collision:

$$(m_A + m_B)v_2 = m_A v_{A1} \quad (4.10.1)$$

There is no change in potential energy when the collision takes place on a level track. Let us explore what happens to the kinetic energy in such collisions. Does the KE increase, decrease, or remain unchanged? The expression for change in KE (final minus initial value) is given by

$$\Delta(\text{KE}) = \frac{1}{2}(m_A + m_B)v_2^2 - \frac{1}{2}m_A v_{A1}^2 \quad (4.10.2)$$

Question 4.10.1 Use Eq. 4.10.1 to obtain an expression for v_2 , and then eliminate v_2 from Eq. 4.10.2, obtaining

$$\Delta(\text{KE}) = - \left(\frac{m_B}{m_A + m_B} \right) \left(\frac{1}{2}m_A v_{A1}^2 \right) \quad (4.10.3)$$

and finally

$$\frac{\Delta(\text{KE})}{\text{Initial KE}} = - \left(\frac{m_B}{m_A + m_B} \right) \quad (4.10.4)$$

Now interpret Eq. 4.10.4 very carefully and completely by addressing the following questions:

- (1) Why bother obtaining the final fraction? Why not stop at Eq. 4.10.3?

- (2) What is the meaning of the minus sign? What happens to KE in this type of collision?
- (3) What fraction of the initial KE is lost when the bodies have equal masses? What happens to this fraction when m_A is made larger and larger relative to m_B ? when m_A is made smaller and smaller relative to m_B ? Note that the vector momentum is *always* conserved even though in the latter case almost all of the KE is lost. (It turns out that such disappearance of KE is always accompanied by an increase in temperature of the colliding objects.)
- (4) What happens to virtually all of the KE of a small bullet that strikes, and lodges in, a massive object?

4.11 USING THE NEW VOCABULARY

When we are learning a new language, the only way to master it is to practice using it. Even though we have been using English words in the preceding development, we have nevertheless been generating a new vocabulary, which is, in essence, a new language. To fully understand the energy concepts, it is necessary to practice using the new language in describing familiar phenomena. Let us now initiate such practice. We start by giving an example. (This exercise is meant to be purely verbal; numerical calculations are not needed.)

Example Describe, in the language of “systems” and “energy transformations,” the entire sequence of changes that takes place when we throw a ball vertically upward and it rises and falls back to lie on the ground. Treat air resistance as negligible, i.e., neglect those dissipative frictional effects that are associated with motion through the air.

Response Let us define the system to be described as consisting of the ball and the earth (*not* including our own body). We take the reference level for zero potential energy at the elevation y_1 above the ground at which our hand begins the act of throwing (i.e., we take $y = 0$ at the surface of the ground). We now describe the transformations taking place in this system:

- (a) Starting with our hand in its lowest position y_1 , our hand exerts an upward force on the ball, displacing the ball vertically upward and doing work on the system. [The energy corresponding to this work comes from complex chemical transformations taking place within our body and in our muscles, but our body is not part of the system under consideration, and this description need not be part of the present story. It is mentioned only in order to help the reader recognize that still another system (our body) is involved and that other energy transformations are occurring outside the system under consideration.]

- (b) The work done on the system (ball-earth) by the force being applied by our hand is converted (partly) into the change in PE stored in the system in elevating the ball from its initial level y_1 to the level y_2 at which our hand loses contact with the ball and partly into the KE imparted to the ball over the same interval.
- (c) After the ball loses contact with our hand, it continues moving upward. As it does so, the KE it possessed at level y_2 decreases, being converted into further increase in the PE of the system. The ball is at the top of its flight at the level y_3 at the instant the KE and the velocity are both zero. The KE has *all* been converted into PE of the system. At this point the work originally done on the system by the force exerted by our hand has been entirely converted into the change of PE of the system. (If we were not neglecting the frictional effect of the air, we would recognize that some decrease in KE had to be ascribed to this dissipative effect, and that the height y_3 attained, and therefore the total change in PE, would not be quite as great as we are now imagining.)
- (d) As the ball falls back down from level y_3 , the PE of the system decreases, and the change is converted into KE of the ball. At the instant of return to level y_1 , the net change in PE of the system is zero, and the KE possessed by the ball is equal to the original total amount of work that was done by the force exerted by our hand.
- (e) The ball continues to fall from level y_1 to the ground (at which we took $y = 0$.) The PE of the system becomes negative by the amount of change corresponding to the drop in elevation between y_1 and the ground, and the KE of the ball increases further by this amount. The total KE of the ball on striking the ground is equal to the original work done by our hand plus the (relatively small) amount of PE transformed between level y_1 and the ground.
- (f) The ball undergoes a perfectly inelastic collision with the ground, and, in this instance, *all* of its KE is dissipated. (It is converted into thermal energy, raising the temperature of both the ball and the region of the ground where the ball landed. We shall say more about thermal energy in forthcoming sections. Note that we do not say that “heat has been transferred to the bodies.” We reserve this terminology to describe cases in which there is a thermal interaction between two systems in contact at different temperatures. In this instance there is no such interaction. The temperature increase makes it *look* as though heat were transferred, but there is actually no such transfer.)

Having treated the above example in detail, we propose the following exercises. You would do well to write out your descriptions and to check and

argue about them with fellow students. You will find that learning to use the language clearly and precisely is a very worthwhile investment of time and effort. It will be enormously helpful in leading to understanding of what is being presented in textbooks and lectures as well as in the solution of problems, and it will lead you to see more deeply into all sorts of changes going on in the world around you.

Question 4.11.1 Describe the following events in the language of “systems” and “energy transformations.” Be sure to define your systems clearly and explicitly in each instance, and follow the pattern of description exhibited in the preceding example.

- (a) A cart is given a push along a level floor. After coasting a short distance along the floor, the cart coasts up an inclined plane and then coasts back to the original horizontal level. (Neglect friction.)
- (b) A pendulum bob of mass m is suspended from point O on a string of length L . Position 1 is that of the bob at its lowest point with the string oriented vertically. Position 2 is that of the bob when the deflection of the pendulum is 90° , and the string is stretched out horizontally. You let the bob go from position 2 and let it swing freely. Describe, in detail, the motion sequence (positions, velocities, and accelerations) that then results in the absence of appreciable resistance from the surrounding air. Then describe the sequence of events in energy transformation following the pattern established in the preceding example of throwing the ball vertically upward.
- (c) Suppose the block and spring system illustrated in Fig. 4.6.1 is suspended vertically so that the block can oscillate up and down but is initially stationary at its equilibrium position under the new circumstances. You pull the block downward, say 2 cm, and then let it go. Describe the motion that results in the absence of appreciable resistance from the surrounding air. Then describe the sequence of events in the language of energy transformations. (Note that *two* forms of PE are now involved simultaneously, not just one.)
- (d) Consider the case in which two carts, having very soft, flexible spring bumpers, collide with each other on the laboratory table. One cart is initially stationary, while the other approaches it with a non zero initial KE. Describe, in the language of energy transformation, the sequence occurring between the instant of initial contact of the bumpers and the instant at which contact ceases. Sketch a rough graph of what the force versus clock reading history must be like for the force one cart exerts on the other during the interval of contact. Describe the sequence of energy transformations occurring when one steel ball bearing collides with another and you are not able to see any deformations.

It is to be carefully noted that a story in terms of the energy concepts does *not* describe what happens in a system as a function of *time*. It does not describe the motion of a body within the system (such as the oscillation of the puck on the spring) as a function of time. It does not describe the interaction

between two bodies in collision as a function of time. Neither did application of the law of conservation of momentum give us histories of variation with time. Equations that give descriptions of position and velocity of objects in a system as functions of time have to come from application of $\vec{F}_{\text{net}} = m\vec{a}$ and the other Newtonian laws. The energy and momentum concepts, however, together with their conservation relations, give us very powerful insights into the connections between the initial and final *states* of a system even when the detailed history of variation with time might be difficult, or even impossible, to obtain.

4.12 MODELS FOR THE NATURE OF HEAT

Most of the leading natural philosophers of the 17th century (Boyle, Hooke, and Newton among them) held a corpuscular view of the structure of matter (i.e., they visualized material substances as consisting of discrete particles too small to be seen through microscopes) even though there was, at that time, no compelling scientific evidence for such discreteness. (The modern, firmly based, and scientifically substantiated atomic-molecular theory had its beginnings in the work of John Dalton early in the 19th century and was not firmly established until around the middle of the 19th century.) Those who held the corpuscular view of matter, thought of what we have called “transfer of heat” as a process in which the invisible corpuscles were set into more violent agitation and vibration. No one was successful, however, in developing or implementing a theory in such a way as to make fruitful predictions or design convincing experiments.

The point of view changed during the 18th century, which saw a proliferation of theories based on imponderable fluids (material substances without measurable mass.) Chemical change, especially combustion, was visualized as involving the gain or loss of a “principle of inflammability” called “phlogiston.” Electric and magnetic interactions were described in terms of fluids (“effluvia”) that emanated from, and swirled around, bodies that participated in the interactions. Thus the general climate of thought, channeled by the fashionable vocabulary, was one that encouraged a view of heat as still another imponderable fluid being transferred from hotter bodies to colder bodies until equilibrium (no further tendency for transfer) was attained at equalization of temperatures.

This view was, of course, strongly reinforced by what appeared to be a conservation of heat in calorimetric experiments. With equal masses of water at different temperatures ending up at mid temperature when mixed, (i.e., undergoing equal magnitudes of temperature change); with unequal quantities undergoing temperature changes inversely proportional to the masses; and with successful application of the concepts of the “calorie” as a unit of heat transferred and of specific heat of different materials measured in calories/(g)(degree), one could say that the amount of heat leaving one system was equal

to the amount entering the other system until the temperatures became equal. This pointed to something being conserved, and this “something” was visualized as another imponderable fluid. Lavoisier, the great French chemist whose researches on chemical reactions and on oxidation in particular, demolished the phlogiston idea, gave the fluid principle of heat the name “caloric” (from the Latin word for heat, *calor*), and the so-called “caloric theory” was launched on a fruitful history, carrying well into the 19th century. Although this model was eventually abandoned, it was not a simple minded or foolish one. You can learn something important about the nature of scientific thought by following a little of what happened.

The caloric fluid was postulated to have the following properties: (1) It is a material substance that can neither be created nor destroyed; (2) the fluid is elastic, and its particles repel each other but are attracted by the particles of ordinary substances, the magnitude of the attraction being different for different materials; (3) caloric can be either “sensible” (leading to temperature changes) or “latent” (entering or leaving substances during melting or freezing without observable change in temperature). In the former case, it diffuses rapidly among the attracting particles and surrounds each with an “atmosphere” of the fluid; in the latter case caloric fluid combines with the attracting particles in a manner similar to that of chemical combinations.

This model provided very plausible explanations of a number of familiar phenomena besides the conservation of heat in calorimetric experiments. Entrance of the caloric fluid among the particles of a substance would cause the latter to tend to spread farther apart, increasing the pressure on the walls as a gas is heated in a rigid container or causing the expansion of liquids and solids (this part of the picture casually ignores the fact that some substances, e.g., liquid water at temperatures between 0 and 4°C, and ice when melting at 0°C, actually contract as heat is transferred from a higher temperature system.) The particles of ordinary substances were thought to attract each other gravitationally (it actually turns out that these interactions are electromagnetic and that gravitational effects are negligible.) Since the particles must be much closer together in liquids and solids than they are in gases (as evidenced by the differences in density), they must attract each other with much stronger forces in liquids and solids, and the caloric fluid would therefore have less effect in causing expansion of liquids and solids than it would have on gases. The well known marked rise in temperature when a gas is rapidly compressed (as in a tire pump) or a material is rubbed or hammered was accounted for as a squeezing out of caloric fluid from the spaces it occupied among ordinary particles.

These are, of course, essentially qualitative applications of the caloric model. In the hands of Laplace, Fourier, and other mathematically gifted investigators, the model yielded interesting and powerful quantitative predictions and relationships that we shall not have time and space to discuss. The point is that the caloric model came to be widely accepted toward the end

of the 18th century because of its attractive simplicity and many successful applications.

While there was nothing more convincing to replace it, the caloric model continued to be used. Accurate measurement, in calories, of quantities of heat transferred did not, in the final analysis, depend on a “correct” model—all that was needed was a good heuristic⁵ one. The caloric model served the purpose admirably, but it is unlikely that many of its users were as sophisticated as Laplace, Lavoisier, and a few other major figures who sensed its limited, heuristic character. Many investigators must have believed in a literal caloric fluid.

Years later it came to be realized that the correct mathematical results that emerged from apparent use of the caloric model did not really depend on the model at all but stemmed from the fact that correct fundamental mathematical equations had been adopted. Other models, which preserve the same mathematical definitions (for example, quantity of heat Q transferred, expressed in terms of mass, specific heat, and temperature change), yield the same correct results. As we shall see, the caloric model was eventually supplanted by an entirely different picture—one that explains and organizes a far wider range of phenomena, and in which heat transfer is recognized as a form of energy transfer rather than as a material fluid. This theory preserves exactly the same mathematical definition of Q and predicts the same relations that were supposedly derived from the caloric model. Let us see how all this happened; it did not do so without a struggle.

4.13 RUMFORD’S ATTACK ON THE CALORIC MODEL

Despite the wide acceptance of the caloric model, a small number of investigators were dissatisfied with the hypothesis of a material fluid having so many specifically tailored (sometimes described as *ad hoc*) properties. Exemplifying this school of thought was a versatile and brilliant man by the name of Benjamin Thompson (1753-1814). Thompson was born in Woburn, Massachusetts, lived for a time in Concord, New Hampshire (then called Rumford), left America because of royalist sympathies during the Revolution, and made a brilliant career in England and on the Continent as an administrator and natural philosopher. For a time he held (simultaneously) the posts of Minister of War, Minister of Police, Major General, Chamberlain of the Court, and State Councilor under the Elector of Bavaria. When the Elector made Thompson a Count of the Holy Roman Empire, Thompson assumed the name of Count Rumford and henceforth chose to be referred to by that title. Subsequently on taking up residence in England, he founded the Royal Institution, a research center that flourishes to this day. In the first part of the 19th century, it was

⁵An heuristic device is defined as an aid to the imagination, one that facilitates thinking, stimulates investigation, or aids in discovery.

the home base for the major scientific figures Humphrey Davy and Michael Faraday.

Rumford was intensely interested in thermal phenomena. In the course of his investigations, he invented different kinds of stoves, an improved fireplace chimney, steam-heating systems, lamps, and coffee makers. He was the first to appreciate the importance of the phenomenon of convection (the circulation of fluids driven by density differences due to heating and cooling) in the transfer of heat, and he conducted a number of crucial experiments relevant to the caloric picture.

One long-standing dispute centered around whether the caloric fluid was really imponderable, i.e., whether or not it had detectable weight. Some investigators claimed that bodies gained weight when heated. Other experiments purported to show that water gained weight as it froze, and it was even suggested that the caloric fluid had a negative weight. Rumford had a distaste for ad hoc inventions of this kind, and, having access to a very fine balance at the court of the Elector, he set out in 1787 to repeat the weighing experiments. He took great care to avoid spurious effects due to convective air currents, changes in temperature of the balance arms, etc., and, in describing the results of his investigations some years later, concluded that heating and cooling of a substance had no detectable effect on its weight (a fact that we now casually take for granted without thinking that it might require careful confirmation.) This observation clearly supported his doubts concerning the material nature of the caloric fluid.

Rumford then turned his attention to the heat produced by friction, and one of his experiments has become particularly famous. He reported it in 1798 to the Royal Society:

Being engaged lately in superintending the boring of cannon in the workshops of the military arsenal at Munich, I was struck by the very considerable degree of heat that a brass gun acquires in a short time in being bored, and with the still higher temperature (much higher than that of boiling water) of the metallic chips separated from it by the borer.

This was hardly a newly discovered phenomenon, but Rumford sensed that here was an avenue to "further insight into the hidden nature of heat." He showed that the small chips had the same specific heat as the bulk metal and that one could not argue that caloric fluid was set free during boring because the chips had a smaller heat capacity. He further showed that just as much heating took place when a blunt boring tool was used and almost no metal was cut. Rumford regarded as most significant

the . . . remarkable circumstance that the source of heat generated by friction in these experiments appeared to be evidently inexhaustible. It is hardly necessary to add that anything which any

insulated body or system of bodies can continue to furnish without limitation cannot possibly be a material substance; and it appears to me to be extremely difficult, if not impossible, to form any distinct idea of anything capable of being excited and communicated in the manner in which heat was excited and communicated in these experiments, except it be MOTION.

In our modern terminology, we would say that the seemingly inexhaustible effect came from the doing (and dissipation) of work. But Rumford remained puzzled by the origin of motion on the microscopic scale:

. . . although the mechanism of heat [might], in fact, be one of those mysteries of nature which are beyond the reach of human intelligence, this ought by no means to discourage or lessen our ardour . . . to investigate the laws of its operation.

Not having a theory with which to connect the generation of heat with any other factors in his experiments, Rumford was not motivated to make quantitative measurements of the relation between work and heat. He did, however, study the transfer of heat through a vacuum, the behavior of water between 0 and 4°C, and the behavior of mixtures of different liquids, all as part of a systematic attack against the caloric model, and, during the early 1800s, Humphrey Davy, Rumford's protégé at the Royal Institution and one of the most eminent chemists of his time, continued Rumford's line.

Rumford, Davy and their followers did not overthrow the caloric theory. A useful and fruitful theory is not abandoned as soon as it fails to fit one or two stubborn facts. The theory can be modified and adjusted, even though it might creak a bit in the process. Such was the response of scientists who continued to rely on the caloric picture well into the 19th century.

4.14 THE QUANTITATIVE RELATION BETWEEN WORK AND HEAT

The attack begun by Rumford was carried to its completion in the 1830s and 1840s in the groping that took place for a synthesis among interlocking phenomena in mechanics, heat, electricity, and chemistry. The unifying principle turned out to be that of energy. Prominent in the evidence that led to the abandonment of the caloric theory were the quantitative experiments of James Prescott Joule.

James Prescott Joule (1818-1889) was a well-to-do Manchester brewer who devoted himself to science from an early age. Joule was profoundly attracted to experimental science, and he had a vision of a grander conservation law than that of conservation of caloric fluid.

Europe in the 1830s was in full flood of the technological revolution. Industry depended on the steam engine for mechanical power generated by that supplied in burning fuel. Volta had invented the electric battery. Faraday had discovered electromagnetic induction, and primitive electric generators were being used in experiments of all kinds. In this atmosphere, Joule conceived the idea (simultaneously held by other investigators) of a possible quantitative connection between work and heat. Such a connection is not obvious despite the ever present temperature increases accompanying frictional effects; people as able as Rumford and Davy failed to close the gap. Furthermore, if a systematic, quantitative connection between heat and mechanical action does obtain, one can still conceive of any number of possible mathematical relationships, involving different mathematical functions. Joule, with a large supply of the gifted scientist's sixth sense, hit upon what turned out to be the right combination.

In 1840 Joule reported some qualitative experiments on the production of heat by electric current, and in 1843 he presented quantitative results. Using falling weights to turn an electric generator, he immersed in water the conductor carrying the electric current and measured the amount of heat that was transferred to the water. Ignoring all the intervening transformations, he compared, directly, the heat evolved to the work (in excess of friction) necessary to turn the generator. [In our modern vocabulary, we would say that he compared the $\Delta(\text{PE})$ of the lowered weights (less the work dissipated in just turning the generator without generating significant current) with the amount of heat transferred from the current carrying conductor to the surrounding water.] Joule reported 13 measurements with an average result that 838 ft lb of dissipated work had the same effect as the transfer of the amount of heat necessary to raise the temperature of 1 lb of water by 1°F .

Without being diverted by the numerical values, let us be clear on what Joule was finding: He was showing that, when a given amount of work was completely dissipated in a system, a fixed amount of heat was always transferred to surrounding water, and that, if one doubled the amount of work dissipated, one also doubled the amount of heat transferred—all of this regardless of intervening transformations, gravitational, electrical, or otherwise. This was the beginning of the breakthrough on the systematic, but hardly obvious, numerical connection between mechanical energy and thermal effects that, as we mentioned earlier, marked the discovery of the general law of conservation of energy. Without the quantitative connection to thermal phenomena, there was no general, powerful synthesis. In this connection, Joule wrote:

I shall lose no time in repeating and extending these experiments, being satisfied that the grand agents of nature are, by the Creator's fiat, indestructible; and that wherever mechanical force is expended, an exact equivalent of heat is always obtained.

In our modern vocabulary, we no longer use the word “force” in this way. We would say “work dissipated” rather than “force expended.”

Joule went on to do many different experiments: with rapidly expanding air, with water forced through holes in a perforated cylinder, and so forth. He then initiated a long series of increasingly accurate experiments which have become especially famous:

The apparatus . . . consisted of a brass paddle wheel working horizontally in a can of water. Motion could be communicated to this paddle by means of [falling] weights. . . . The paddle moved with great resistance in the can of water, so that the weights (each of four pounds) descended at a slow rate of about one foot per second. The height of the pulleys from the ground was twelve yards, and consequently, when the weights had descended . . . , they had to be wound up again to renew the motion of the paddle. After this operation had been repeated sixteen times, the increase of the temperature of the water was ascertained. . . .

Joule continued his measurements, improving their accuracy and working with different materials, both liquid and solid. By the time he reported an extensive summary of results in 1850, he had worked with the entire array of different transformations mentioned above:

The quantity of heat produced by the friction of bodies, whether solid or liquid, is always proportional to the quantity of [work] expended. The quantity of heat capable of increasing the temperature of one pound of water . . . by 1°F requires for its evolution the expenditure of mechanical [work] represented by the fall of 772 lb through the space of one foot.

Thus Joule explained the source of the “inexhaustible” supply of heat in Rumford’s boring of cannon. As long as work is dissipated, a heating effect is produced. Recognition of heat as another form of energy means that the number which Joule keeps quoting and improving (usually called the “mechanical equivalent of heat”) is simply a conversion factor between different units of measurement (like the conversion of inches to centimeters by the factor of 2.54.) The currently used values for this conversion factor are 778 ft lb for raising the temperature of 1 lb of water by 1°F (called the “Btu” for British thermal unit) or 4.186 J (for “joules,” named in Joule’s honor) for raising the temperature of 1 g of water by 1°C (called the “calorie.”)

It is easy to talk about the number as being simply a conversion factor between different units of measurement for the same thing once one knows that one is indeed dealing with the same thing. The latter insight could not have been achieved, however, had not both heat and work been measured separately in their own units before the insight was achieved.

Joule's experiments, together with the work and thought of a number of contemporaries, led to the formulation of the general principle of conservation of energy that we shall formulate more explicitly in Sect. 4.17. This prodigiously fruitful synthesis of mechanics and theory of heat invited a return to the earlier view of heat as being associated with the motion (kinetic energy) of constituent particles of matter. But the development of the science of energy proceeded on the macroscopic level; it did not require a clear and adequate theory of the structure of matter; it was not necessary to define precisely what was moving and how on the microscopic scale. The essence of the theory lay in mathematical relationships among observable macroscopic properties of matter without relying on elucidation of microscopic phenomena.

This is not to say that a microscopic model was uninteresting or undesirable; such a theory came later (actually developing in stages and not being fully completed until well into the 20th century) and enormously deepened our insights and understanding. You will eventually encounter an analogous situation when we see how Young and Fresnel evolved, early in the 19th century, a highly successful wave model of light without having any clear idea of what was "waving," and how Maxwell ultimately provided a detailed theory of electromagnetism that encompassed light among the phenomena with which it dealt.

Study of the energy relations on the microscopic level involves statistical concepts and theoretical developments you will come to later when you study kinetic theory and, still later, in the science of statistical mechanics. For the time being we shall pursue the subject on the macroscopic level in the spirit in which it was originally developed in the 19th century.

4.15 JOULE'S ENUNCIATION OF THE PRINCIPLE OF CONSERVATION OF ENERGY⁶

In a popular lecture delivered before a meeting of the British Association for the Advancement of Science in Manchester in 1847, Joule sounded the knell of the caloric theory. (In the following quotes you should read "living force" as kinetic energy and "attraction through space" as potential energy):

⁶The sequence we have been describing consists of abridged elements from what is really a rather complex episode in history of science. The reader should note the fragmentary nature of the story and should realize that, although this is useful in helping us gain insight into the development of the concepts, it does not give a fully accurate historical picture. For example, Julius Robert Mayer (1814 - 1878), in Germany, published, in 1842, an initial statement very much along the lines of the following quotation from Joule (somewhat before Joule published his first experimental results) together with an initial statement regarding the conservation principle, in 1843. Mayer's thought, however, was abstract and almost mystical (he was not a skillful experimentalist like Joule), and he provided very little experimental support for his ideas. His prior publication, however, subsequently led to claims and counterclaims over who was entitled to priority for the discovery.

The most prevalent opinion until of late, has been that [heat] is a substance possessing, like all matter, impenetrability and extension. We have however shown that heat can be converted into living force and into attraction through space. It is perfectly clear, therefore, that unless matter can be converted into attraction through space, which is too absurd an idea to be entertained for a moment, the hypothesis of heat being a substance must fall to the ground. Heat must therefore consist of either living force or of attraction through space. . . . I am inclined to believe that both of these hypotheses will be found to hold good—that . . . sensible heat will be found to consist in the living force of the particles of the bodies in which it is induced; whilst in other [instances], particularly in the case of latent heat, the phenomena are produced by the separation of particle from particle, so as to cause them to attract one another through a greater space. [Joule's hypothesis turned out to be correct.]

Joule's sense of exhilaration in the perception of a far-reaching conservation law is evident in the following poetic language from the same lecture:

The motion of the air which we call "wind" arises chiefly from the [high temperature] of the torrid zone compared with the temperature of the temperate and frigid zones. Here we have an instance of heat being converted into the living force of currents of air. These currents of air, in their progress across the sea, lift up its waves and propel the ships; whilst in passing across the land they shake the trees and disturb every blade of grass. The waves by their violent motion, the ships by their passage through a resisting medium, and the trees by the rubbing of their branches together and the friction of their leaves against themselves and the air, each and all of them generate heat equivalent to the diminution of the living force of the air which they occasion. The heat thus restored may again contribute to raise fresh currents of air; and thus the phenomena may be repeated in endless succession and variety.

When we consider our frames, "fearfully and wonderfully made," we observe in the motion of our limbs a continual conversion of heat into living force, which may be either converted back again into heat or employed in producing an attraction through space, as when a man ascends a mountain. Indeed the phenomena of nature, whether mechanical, chemical or vital, consist almost entirely in a continual conversion of attraction through space, living force, and heat into one another. Thus it is that order is maintained in the universe—nothing is deranged, nothing ever lost, but the entire machinery, complicated as it is, works smoothly and harmoniously. And though, as in the awful vision of Ezekiel, "wheel may

be in middle of wheel," and everything may appear complicated and involved in the apparent confusion and intricacy of an endless variety of causes, effects, conversions, and arrangements, yet is the most perfect regularity preserved. . . .

Thus we see, along with Joule, that the losses of mechanical energy that we observed in the presence of frictional effects (the effect of air resistance on a moving ball, the effect of friction when a block slides along the floor, the losses in inelastic or partly elastic collisions) are not actually losses of energy. Mechanical energy alone is *not* conserved in the presence of friction; what is conserved is the *combination* of mechanical and thermal energy.

4.16 EXTENDED BODIES AND SYSTEMS AS OPPOSED TO POINT MASSES

In our first probing for possible conservation relations, we started with the simplest available situations, dealing with bodies such as simple blocks of material that could be treated, from the standpoint of $\vec{F}_{\text{net}} = m\vec{a}$, as single particles concentrated at the center of mass. This led us to the various useful relations we obtained and interpreted in Sects. 4.3 to 4.6 and to the initial hints about energy conservation. Our macroscopic blocks are, however, not simply point masses, and treating them as such is a useful fiction only up to a certain point. We are now at the point where this device can trip us up if we are not careful. Let us look ahead at what might be involved.

Where we first get into trouble with the point-mass picture is in connection with transfer of heat. If heat, on being transferred to an object, is, on the microscopic scale, preserved in the form of kinetic energy of atoms and molecules and in potential energy of their interactions with each other, this "thermal energy," as it is called, is distributed over the entire body of the block and cannot be thought of as concentrated at the center of mass point—even if we do not try to describe the microscopic kinetic and potential energies in detail. We are compelled to think of heat as being converted into internal thermal energy distributed over the entire extended object. Furthermore, we cannot very well visualize and deal with friction between point masses. Friction involves the contact and rubbing of solid surfaces against each other (or the rubbing of layers of fluid against each other or against solids). We now know the rubbing effects between solids, for example, to involve very intricate interactions in which the surfaces of the materials become "welded" together and "un-welded" at numerous points; atoms or molecules from one material actually migrate and interpenetrate the structure of the other material. Thus the interaction intrinsically involves "contact" surfaces of extended objects and cannot be dealt with as though it were taking place between point masses.

When we push a box along the floor with uniform velocity, the conversion of the work we are doing into thermal energy takes place all over the surfaces

in contact and has nothing to do with the centers of mass of the rubbing objects. Thermal energy is acquired not only by the box but also by the floor: the effects cannot be confined to just one of the bodies in contact; both bodies (the box and the floor) exhibit an increase in temperature. Note that we have just described the effects taking place in this system (box and floor) as involving a conversion of work (done by us on the system) directly into thermal energy of the objects in contact and not as a transfer of heat to the system. We reserve the term “transfer of heat” for situations in which thermal energy is transferred from a system at higher temperature to a system at lower temperature without any displacements that involve the doing of work by one system on the other. Sticking to this terminology, we must say that no heat was transferred from any outside system to the box-floor system since there is no contact with a higher temperature body. The temperature increase that occurs takes place through the direct conversion of mechanical work into thermal energy within the system without any transfer of heat.

Other situations in which we obviously get into difficulty with the point-mass simplification are those in which objects or systems we are considering undergo deformation, as in cases such as compressing a spring (recall that the formula that gives us conserved values for the work done and energy stored in compressing a spring involves the displacement of the end of the spring and not the displacement of its center of mass) or compressing a gas in a cylinder, or cases of deformation of our own body when we walk, or run, or jump vertically upward, or push ourselves horizontally.

In the light of these various examples, it is clear that we must refine and amplify our statement of the energy conservation concept so as to embrace extended systems, not just point masses. This is what we now proceed to do.

4.17 HEAT, WORK, AND CHANGE OF STATE: THE FIRST LAW OF THERMODYNAMICS

We now put together the statement (or law) that turns out to describe what we observe happening in natural phenomena involving energy transformations and the associated changes of state of bodies or systems. This law cannot be derived from some absolute, fundamental principle any more than $\vec{F}_{\text{net}} = m\vec{a}$ or the law of gravitation can be derived. Such laws are initially put forth as guesses or conjectures based on the hints and restricted numerical relations that have been assembled, i.e., by inductive reasoning. They then have to be tested to see whether or not they “work,” i.e., describe and predict natural events and relationships correctly. Only after one finds no failures in many applications does one begin to accept and trust the law. (We went through such a process with the law of conservation of momentum.)

Let us start by reviewing the meaning of “work” and “heat.” We give the name “work” and the symbol W to the calculation of energy that is transferred

to or from a system when a force acts through a corresponding displacement. Such transfer involves an interaction between systems or objects. We shall take the work quantity to be positive when work is put into (or done on) the system under consideration and negative when the system does work on its surroundings. (This is not a universal convention; when you study the science of thermodynamics, you will find that the sign convention adopted is frequently opposite to the one we have just enunciated.) Note that we are restricting the meaning of the term “work” to that which goes into and out of the *system* we are considering. Energy transformations that take place *inside* the chosen system are *not* described in terms of work unless we separate our system into sub-systems. Only then might we talk about the work one sub-system does on another.

We give the name “transfer of heat” and the symbol Q to that form of energy (characterized as “thermal”) that is transferred from a higher temperature system to a lower temperature system when the two systems are in “thermal contact” in the absence of the interaction that we have called “work.” As described in connection with the discussion of frictional effects in the preceding section, Q describes thermal interaction of the system under consideration with some other system, and, if there is no such interaction taking place, no heat is transferred and $Q = 0$ *even though temperature changes may be taking place within the system being considered*. We take Q to be positive when heat is transferred *into* a system under consideration, and we take Q to be negative when heat is transferred *out* of the system.

Note that, in the light of these definitions of the terms “work” and “heat” as quantities of energy being *transferred*, it is incorrect and inappropriate to talk about either work or heat as being *in* a system, or as though they were properties of a system. They are *not* properties. One should not use phrases such as “heat in the system” or “heat of the system.” The terms “work” and “heat” are, in modern usage, restricted to refer only to energy being transferred from one system to another. On the other hand we can correctly talk about *properties* of a system as being *in* or *of* the system. Pressure, temperature, volume, mass, density, kinetic energy, potential energy are all intrinsic properties that one can correctly speak of as being *in* or *of* a system.

The preceding statements about W and Q are descriptions as to how to calculate certain quantities that appear to be related but are not, in themselves, laws of nature. The law of nature we are getting at has to do with the connection between the quantities we have defined and the *changes in state* of bodies or systems that are subjected to the energy transfers represented by W and Q . Following is the way experience shows that nature operates:

- (a) If we transfer heat *into* a system through contact with a system at higher temperature (without allowing any transfer of work), we change the state of the system in some way, the simplest case being an increase in temperature and pressure, but other, more complex, changes are possible,

such as melting, dissolving, chemical or electrical changes, and so forth. If we then take out an equal amount of work (without further transfer of heat), the system returns to its initial state, i.e., to the same temperature and pressure at which we started. The same occurs if we put in a given amount of work and then take out an equal amount of heat: We return the system to its initial state. In symbols, what we are saying is that, if $Q + W = 0$, we do not change the state of the system. We change the state only if $Q + W \neq 0$. This is not something “derived” nor is it at all obvious. This is the way heat and work transfers are found to affect the state of systems. (Note how this statement connects with the impossibility of the “getting something for nothing” perpetual motion devices visualized in Question 4.1.1 and in the discussion of pushing an object up the inclined plane in Sect. 4.8. In such devices we would be extracting an inexhaustible supply of work or kinetic energy with no *net* change in the state of the system after endless repetition of the cyclic process.) If we put in work alone or heat alone, we always change the state of the system.

- (b) If the quantities of work and heat transferred do not add up to zero, there is a net change in the state of the system, and a net amount of energy has either been added or taken out. This leads us to think about the changes in state of the system as indicating changes in energy within the system: changes such as in kinetic energy of moving parts, or potential energy stored or released in compressions or expansions, or thermal energy associated with increase or decrease in temperature. To such energy quantities within the system, we give the name “internal energy” and the symbol E . Changes in internal energy will, of course, be denoted by the symbol ΔE . In the case that $Q + W = 0$, it follows that $\Delta E = 0$.

- (c) When $Q + W \neq 0$, it turns out that the following relation always holds⁷:

$$\Delta E = Q + W \quad (4.17.1)$$

The law expressed in Eq. 4.17.1 cannot be derived. It was guessed at originally in the years following the work of Joule and his contemporaries and has never been found to fail. It is the broad, general assertion of conservation of all forms of energy. As investigators tackled increasingly complex systems (such as those including chemical, electrical, and magnetic phenomena) the concept of internal energy had to be extended to include quantities of energy associated with such effects. It was a matter of discovering the “right formulas”

⁷When the opposite convention for the algebraic sign of W is used, i.e., when W is taken as positive if work is done *by* the system *on* its surroundings and energy leaves the system, this statement is written in the form $\Delta E = Q - W$. You may encounter this form in the study of thermodynamics. There is, however, no need for worrying about the difference now.

for calculating these effects (as Feynman says in the remark quoted in Sect. 4.1.) Such formulas have always been found, and Eq. 4.17.1 has always held up. The consequence is that it is regarded as one of the most basic, general, and far reaching laws, or expressions of regularity, in nature. It is called the “first law of thermodynamics,” and we shall use that name from here on, abbreviating it, for convenience, with the acronym FLT.

Question 4.17.1 Choose some familiar events in everyday experience and describe what takes place in terms of Q , W , and ΔE .

- (a) For example, in throwing a ball vertically upward in the absence of air friction, we first put work W into the ball-earth system, $Q = 0$ since there is no transfer of heat, and ΔE consists of an increase in kinetic energy of the ball and an increase in the PE of the system. Now continue by describing what happens as the ball rises to the top of its flight.
- (b) Describe in a similar way what happens when we heat a pot of water on a stove; when we push on a rigid wall; when we compress air in a bicycle pump very rapidly; when we compress the air very, very slowly; when a box sliding on the floor comes to a stop under the influence of friction; when two carts undergo perfectly elastic collision on an air track; when the two carts undergo perfectly inelastic collision; when a fire extinguisher containing compressed carbon dioxide gas is opened. In each instance be very careful to define the system you will discuss. Remember that processes taking place at constant pressure usually involve expansion or contraction and therefore exchange of W with the surroundings.
- (c) Analyze some instances of your own choice and invention.

4.18 THE VARIETIES OF INTERNAL ENERGY

In preparation for further examples and discussion, let us list the different kinds of internal energy change with which we shall be concerned (even if we do not develop formulas for all of them) and specify the symbols we shall use to represent each one:

- (a) Internal thermal energy change (associated with change in temperature and/or melting or freezing): ΔE_{therm}
- (b) Internal chemical energy change: ΔE_{chem}
- (c) Internal kinetic energy change: ΔE_{kin} or $\Delta(\text{KE})$
 - (1) Internal KE changes would have sub-categories such as change in translational KE, $\Delta[(1/2)mv^2]$, denoted by $\Delta E_{\text{kin trans}}$ and change in rotational KE, denoted by $\Delta E_{\text{kin rot}}$ ⁸

⁸Rotational KE is very much like translational KE, except that the calculation is a bit more complex. In a rotating wheel, for example, only particles at the same radius have

- (d) Internal potential energy changes: ΔE_{pot} or $\Delta(PE)$
- (1) Internal PE changes would have sub-categories such as gravitational, denoted by $\Delta E_{\text{pot grav}}$; spring-like, denoted by $\Delta E_{\text{pot spr}}$; electrical, denoted by $\Delta E_{\text{pot el}}$; and so forth.
 - (2) Examples: In the simple case in which we change the vertical position of an object of mass m near the surface of the earth by the amount Δy , $\Delta E_{\text{pot grav}} = mg\Delta y$ for the earth-object system. In the simple case of compressing a spring within the system under consideration, $\Delta E_{\text{pot spr}} = \Delta[(1/2)kx^2]$, where, as before, x denotes the displacement of the end of the spring from its relaxed position.
- (e) Miscellaneous internal energy changes ΔE_{misc} , encompassing emission or absorption of sound or light, or other messy interactions for which we have not developed formulas at this point.

The general symbol ΔE in the FLT (Eq. 4.17.1) now stands for the algebraic sum of the different internal energy changes specified in the above list. [We use the word “algebraic” because some of the changes may be increases (therefore positive) while others may be decreases (therefore negative).] We then have:

$$\Delta E \equiv \Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin}} + \Delta E_{\text{pot}} + \Delta E_{\text{misc}} + \dots = Q + W \quad (4.18.1)$$

which is simply an expanded version of Eq. 4.17.1.

Let us illustrate the meaning of Eq. 4.18.1 by applying it to some very simple cases:

- (a) If we transfer heat in or out of the system by bringing it in contact with a higher or lower temperature system and, at the same time, allow no work to be transferred ($W = 0$), Eq. 4.18.1 gives:

$$\Delta E = Q \quad (4.18.2)$$

and if there are no chemical, KE, or PE changes (as in the case of simply heating a beaker of water), Eq. 4.18.2 reduces further to

$$\Delta E_{\text{therm}} = Q \quad (4.18.3)$$

In other words, in such simple cases, all the heat transferred into or out of the system goes into increasing or decreasing thermal internal energy as manifested by the change in temperature and/or change in phase. We have said this before in words, but now we verify that the FLT says it in symbols.

the same linear speed v . Those at larger radii have larger values of v , and those at smaller radii have smaller values. To find the total KE of a rotating wheel, it is necessary to add up the different $(1/2)(\Delta m)v^2$ values for each small chunk of mass Δm at different radii. This involves an integration over the entire wheel. You will come to this later in discussing rotational dynamics. We do not need the formula now.

- (b) If we do work on the system or allow the system to do work on its surroundings while it is thermally insulated and no heat is transferred in or out (i.e., if $Q = 0$), we have, from Eq. 4.18.1:

$$\Delta E = W \quad (4.18.4)$$

Such processes (with $Q = 0$) are called “adiabatic.” In the very simple case of accelerating a particle horizontally in the absence of friction with no transfer of heat, we have such an adiabatic process, and Eq. 4.18.4 reduces to

$$\Delta E_{\text{kin trans}} = W \quad (4.18.5)$$

which is a way of saying Eq. 4.7.2 in our new language.

Question 4.18.1 Take the episodes you considered in Question 4.17.1 and set them up in terms of the symbols of the FLT in exactly the same manner as in the immediately preceding illustrations.

4.19 DEALING WITH EXTENDED SYSTEMS

Systems or bodies that can be dealt with as single particles (as in the cases analyzed in Sects. 4.3 to 4.8) yield extremely limited insight into energy transformations and the restrictions imposed by the FLT. In these sections we were able to say virtually nothing significant about heat transfer, about frictional effects, about deformable systems such as our own body when running or jumping, or about materials that are being compressed (such as gas in a cylinder or material deep within the earth), and so forth. Most phenomena of real interest involve extended systems rather than bodies that can be treated as single particles. We must re-examine our governing equations from the point of view of applicability to extended systems.

The first law of thermodynamics (FLT) as stated in Eq. 4.18.1 is perfectly general and applies to *all* systems. It requires no modification. The relation we obtained, however, by integrating Newton’s second law over the displacement of a *particle* on which an unbalanced force is acting, namely

$$\int_{s_1}^{s_2} \vec{F}_{\text{net}}(s) \cdot d\vec{s} = \Delta \left(\frac{1}{2} m v^2 \right) \quad (4.19.1)$$

requires a very important alteration in order to use it in dealing with extended bodies or systems.

When we deal with an extended body or system rather than a single particle, Newton’s second law takes the form

$$\vec{F}_{\text{net}} = m \vec{a}_{\text{cm}} \quad (4.19.2)$$

in which the subscript cm reminds us explicitly that the acceleration imparted by the net force is that of the center of mass of the system. Various parts of the system may not share the same acceleration if, for example, the system is being deformed by internal effects or by the action of the forces on its periphery. (When we stand on roller skates and push ourselves away from a wall, for example, our hands and our arms, while flexing, do not have the same acceleration as the center of mass of our body as a whole.)

When we integrate both sides of Eq. 4.19.2 with respect to position s , we must keep in mind that the position number we are talking about is that of the center of mass only and that we can no longer represent the entire extended system as a particle concentrated at this position. On integration of Eq. 4.19.2, following the same mathematical procedures as in earlier sections, we obtain the same form as in Eq. 4.19.1 except for the cm subscript:

$$\int_{s_1}^{s_2} \vec{F}_{\text{net}}(s) \cdot d\vec{s}_{\text{cm}} = \Delta \left(\frac{1}{2} m v_{\text{cm}}^2 \right) \quad (4.19.3)$$

and, if the net force is constant and in the same direction as the cm displacement, Eq. 4.19.3 reduces to the simpler form

$$F_{\text{net}}(\Delta s_{\text{cm}}) = \Delta \left(\frac{1}{2} m v_{\text{cm}}^2 \right) \quad (4.19.4)$$

Eqs. 4.19.3 and 4.19.4 look very much like those applying to the single particle, but there turns out to be a very important physical difference: The velocity on the right hand side, v_{cm} , does *not* apply to every object or particle in the extended system but applies *only* to the center of mass point. Similarly, the displacement, Δs_{cm} , on the left hand side represents the displacement of the center of mass point, and it is not necessarily the displacement of the various forces, applied around the periphery of the system and comprising the net force. When you push yourself away from the wall, for example, the force exerted on you by the wall undergoes no displacement and does zero work; yet your cm is displaced.

The consequence of this change in meaning from our earlier single particle case is that the left sides of Eqs. 4.19.3 and 4.19.4 do not necessarily represent an amount of work done on the system by its surroundings, i.e., the left hand side of these equations is not necessarily the same as the quantity W in Eq. 4.18.1. We shall develop a better understanding of these distinctions by applying the governing equations to a number of now familiar physical situations.

4.20 THE BLOCK AND SPRING WITHOUT FRICTION

Let us start by re-examining the situation we analyzed in Sect. 4.6 (the spring and the frictionless puck undergoing horizontal displacements as reproduced

in Fig. 4.20.1) where we dealt with the particle and spring separately. In that analysis, we focussed all our attention on the puck as a particle and said nothing about the spring except that it stored and released PE. We now consider the frictionless puck and spring together as a system subject to the FLT as previously expressed in Eq. 4.18.1 and as now shown again, for convenience, in Eq. 4.20.1:

$$\Delta E \equiv \Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin}} + \Delta E_{\text{pot}} + \Delta E_{\text{misc}} + \dots = Q + W \quad (4.20.1)$$

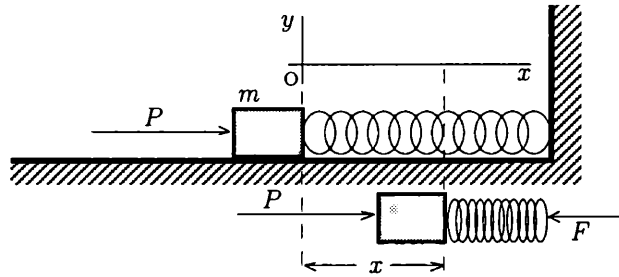


Figure 4.20.1 Block and spring without friction, as in Fig. 4.6.1. Rather than taking the block alone as the system under consideration, the block and spring are now treated as a *combined* system. The mass of the spring is *not* considered negligible.

Since the force F exerted by the wall on the spring undergoes no displacement, it does no work on the system; in the terminology developed in Sect. 4.7, F is a zero work force. Thus work W is put into the system *only* by the external force P exerted by us. There is no transfer of heat from a higher temperature system, so $Q = 0$. If we displace the left-hand end of the spring from its relaxed position at $x = 0$ to some position $x = x_1$, we do a specific amount of work, numerical value denoted by W_1 , on the puck-spring system. W_1 is then given by the integral

$$W_1 = \int_0^{x_1} P(x) dx \quad (4.20.2)$$

We assume the puck starts from rest at $x = 0$ at the instant we apply the force P . We assume the spring obeys Hooke's law and denote the spring constant by k . We recognize that the work done on the system is transformed into the internal energy forms of PE and KE. All other internal energy changes are zero. Thus, for the state of the system at the instant the puck reaches the position x_1 , Eq. 4.20.1 gives:

$$\Delta E_{\text{kin trans}} + \frac{1}{2} k x_1^2 = W_1 \quad (4.20.3)$$

where the translational kinetic energy term includes both the kinetic energy of the puck $(1/2)mv_1^2$ at position x_1 plus the kinetic energy distributed among

the coils of the spring, which, at this point, we would be hard pressed to calculate but which we can at least recognize in principle. (In our previous analysis in Sect. 4.6, we avoided this difficulty by treating the spring as having negligible mass.)

Now, at the instant we have put the amount of work W_1 into the system and the puck is at position x_1 , we remove force P , adding no more energy, and leaving the puck free to keep moving as the conditions allow. What will happen from here on? Let us go back to Eq. 4.8.9, which we reproduce for convenience as Eq. 4.20.4:

$$\text{PE} + \text{KE} = H \quad (4.20.4)$$

Notice the connection between Eqs. 4.20.3 and 4.20.4. W_1 corresponds to the total initial energy H . Eq. 4.20.4, applied to our system, says that, after we have put the amount of work W_1 into the system and let the system continue its subsequent history of change (in the absence of friction), the internal energy keeps transforming back and forth between potential and kinetic, the instantaneous sum of the two forms being always equal to W_1 . Notice also, however, that the work-KE theorem in the form of Eq. 4.19.3 is not helpful in description of energy transformations in the block-spring system since we do not know either the location or displacement of the center of mass.

Question 4.20.1 Analyzing and interpreting Eq. 4.20.4 as applied to our system, argue that, since maximum displacement must occur at the instant $\text{KE} = 0$, the maximum displacement of the end of the spring x_{\max} must be given by

$$x_{\max} = \pm \sqrt{\frac{2W_1}{k}} \quad (4.20.5)$$

Interpret Eq. 4.20.5. What is the significance of the negative sign? Does it have physical meaning or is it uninterpretable nonsense?

Question 4.20.2 Let us go back to the idealized case in which not only is the friction assumed to be zero but the spring is taken to have negligible mass. Argue that, after W_1 has been put into the system and force P drops to zero, Eq. 4.20.3 takes the form

$$\frac{1}{2}mv^2 + \frac{1}{2}kx^2 = W_1 \quad (4.20.6)$$

Interpret what Eq. 4.20.6 says about the motion of the puck: At what positions is the displacement a maximum? What is the value of v at these positions? At what positions is the velocity at a maximum? What is the expression for v_{\max} ?

Question 4.20.3 Explain in your own words the difficulties you encounter if you try to analyze the combined puck-spring system by trying to use the cm Eq. 4.19.3 instead of Eq. 4.20.1.

4.21 PERSON JUMPING VERTICALLY UPWARD

Now let us analyze the somewhat more complex case in which a person (with mass m) jumps vertically upward from a standing position: Where does the energy involved in this action come from and what transformations take place?

Question 4.21.1 Imagine that you yourself are doing the jumping. Draw a force diagram of your body, showing the forces acting on you while you are still in contact with the ground. Draw a force diagram for the ground in the region near your feet. Label the magnitude of the gravitational force the earth exerts on you (your weight) mg . Label the magnitude of the normal force exerted on you by the ground N .

First let us apply the cm Eq. 4.19.4 and interpret the information it yields. (We shall use the constant force form 4.19.4 rather than the more general integral form 4.19.3 by simplifying the situation through electing to talk about the *average* force \bar{N} exerted by the ground on you rather than trying to consider the variation of N during the actual history of contact with the ground.) Eq. 4.19.4 is reproduced as Eq. 4.21.1 for convenience.

$$F_{\text{net}}(\Delta s_{\text{cm}}) = \Delta \left(\frac{1}{2}mv^2 \right) \quad (4.21.1)$$

In the act of jumping, your center of mass starts from rest, acquiring a change in elevation Δh_{cm} and a final velocity $v_{\text{cm}f}$ at the instant your feet leave the ground.

Question 4.21.2 Taking the positive direction upward, argue that, while your feet are still in contact with the ground, the net force acting on you must be $\bar{N} - mg$. Then show, from Eq. (4.21.1), that

$$(\bar{N} - mg)\Delta h_{\text{cm}} = \frac{1}{2}mv_{\text{cm}f}^2 \quad (4.21.2)$$

and that

$$\bar{N}\Delta h_{\text{cm}} = mg\Delta h_{\text{cm}} + \frac{1}{2}mv_{\text{cm}f}^2 \quad (4.21.3)$$

Eq. 4.21.2 looks very much like a work-kinetic energy equation, but it is very important to understand that it is *not*. The normal force N undergoes zero displacement and does zero work on you in the act of jumping! If this were a true energy equation it would be saying that the energy which appears as your KE and as the increase in PE of the earth-you system came from the external action exerted on you by the ground. The ground, however, certainly does not supply energy to the system consisting of you and the earth. The energy changes we observe come from the internal chemical energy changes taking place within your body and not from the outside. If the necessary

energy came from the ground rather than from within you, you could go on jumping forever without any diminution of your ability to do so!

This observation does not mean that Eq. 4.21.2 is in any way incorrect. It is derived from Newton's second law, and is a perfectly correct *dynamical* equation, relating the force N to your change in elevation and the velocity imparted to your center of mass. But it is not a true energy equation, even though the quantities on the right hand side are indeed the PE and KE changes imparted to the jumper-earth system. This problem of interpretation arises because the product of an external force and the displacement of the center of mass of the body on which it acts is not, in general, an amount of work done on the system in question. This product is an amount of work if the system is a simple particle since the external force will then automatically be displaced with the particle, but, if the system is extended and deformable as in the case of the jumper, the product of force and displacement of center of mass is not necessarily an amount of energy transferred to the system. (Some authors have suggested calling such a quantity "pseudo work" to distinguish it from "real" work, but this terminology has yet to be widely accepted. For the time being you can select your own name if you like.) \bar{N} , like F in Fig. 4.20.1, is a zero work force since it undergoes zero displacement even though the cm of the body is displaced.

What we are now seeing illustrated, is the profound difference between Newton's second law (and relations derived from it) on the one hand and the first law of thermodynamics on the other. The FLT is an independent statement about order in nature and is not derivable from Newton's second law. In order to apply the FLT, we must calculate work in accordance with the definition [Eq. 4.7.1]:

$$W \equiv \int_{s_1}^{s_2} \vec{P}(s) \cdot d\vec{s} \quad (4.21.4)$$

where the displacement over which we integrate is that of the external force itself and not that of the center of mass of the object on which it acts. In some instances this displacement may be identical with that of the cm (as it always is in the case of a particle), but in many important instances it is not. In the case of the jumper, the displacement of the normal force acting on the jumper is zero, while the displacement of the center of mass is clearly not zero. This difference stems from the fact that the jumper's body is deformable, and the cm is displaced without displacement of the force \bar{N} exerted by the ground.

Recall that the object of our investigation is to *discover* how to calculate the numbers that are preserved in interactions resulting in changes of state. We must separate the calculations that "work" in this fashion from those that do not.

Let us now do a proper energy analysis of the earth-jumper system (not including the surrounding air) by applying the FLT:

$$\Delta E \equiv \Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin}} + \Delta E_{\text{pot}} + \Delta E_{\text{misc}} + \dots = Q + W \quad (4.21.5)$$

Since no external work is done on or by the system, \bar{N} being a zero work force, $W = 0$. Since the system might exchange heat with higher or lower temperature surroundings, we leave the term Q in our equation, realizing that no heat is put into the system from the surroundings but that the jumper might lose some heat to the air as he or she warms up from the exercise. (Under these circumstances Q would have a negative value.) The relevant internal energy changes are: chemical (the origin of the biological effects in the action of the muscles), thermal (the warming up of the entire body from the exercise), kinetic translational $[(1/2)mv_{\text{cmf}}^2]$, kinetic rotational (flailing of arms, if this occurs), and gravitational potential ($mg\Delta h_{\text{cm}}$). Eq. 4.21.5 then becomes:

$$\Delta E_{\text{chem}} + \Delta E_{\text{therm}} + \left(\frac{1}{2}mv_{\text{cmf}}^2\right) + \Delta E_{\text{kin rot}} + mg\Delta h_{\text{cm}} = Q \quad (4.21.6)$$

To simplify matters, let us neglect the rotation of the arms and thermal effects within the body. With negligible thermal effects, there is no transfer of heat to the surroundings owing to the act of jumping. (There will be just the normal heat transfer that is taking place all the time because our body maintains a higher temperature than our surroundings. This heat transfer is not part of the present problem.) Therefore, for the present problem, $Q = 0$. Eq. 4.21.6 then reduces to

$$\Delta E_{\text{chem}} = -\frac{1}{2}mv_{\text{cmf}}^2 - mg\Delta h_{\text{cm}} \quad (4.21.7)$$

This equation tells us that the source of upward KE of the jumper and of the increase in PE of the jumper-earth system resides in the *decrease* of internal chemical energy (stored as a kind of potential energy) within the body of the jumper. This chemical energy is transformed through work transferred between sub-systems within the body by forces and corresponding displacements of muscles and body parts that we cannot describe in any detail. The cm equation (4.21.3) shows that the right hand side of Eq. 4.21.7 happens to be numerically equal in magnitude to $\bar{N}\Delta h_{\text{cm}}$, despite the fact that \bar{N} is not a work doing force and supplies no energy to the earth-jumper system.

The latter point merits emphasis and re-statement: The force \bar{N} , acting at the feet of the jumper, does impart *acceleration* to the center of mass of the jumper, but despite the fact that it imparts acceleration, it does zero work and supplies zero energy to the system. In other words, we need *both* Eqs. 4.21.3 and 4.21.7 to tell us the whole story of acceleration and energy transformations. Eq. 4.21.3 is frequently *incorrectly* interpreted as an energy relation. It is *not* an energy relation, and such interpretation should be carefully avoided.

An inverse kind of illustration emerges when we consider what happens if we compress a spring by pushing it actively with equal and opposite forces at each end. (In this case we do not displace its center of mass as we did in Fig. 4.20.1 where we pushed on only one end and had the rigid wall exert

a passive force with no displacement at the other.) In this case the cm of the system (spring) is neither accelerated nor displaced and the work-kinetic energy theorem (Eq. 4.19.4) tell us nothing and is, in fact, misleading. Since the two forces acting on the spring are equal and opposite, and since the cm is not displaced, all the terms in Eq. 4.19.4 are zero, and, if interpreted as an energy relation, this implies that no energy is put into the system.

This is, of course, not the case at all. Each of the two forces at the opposite ends *does* do work on the system, and the total amount of work done is stored as PE in the compression of the spring without any displacement of the cm. Thus, it is not necessary for the cm of a system to be accelerated or displaced for work to be done on the system and transformed into some form of internal energy. On the other hand, as illustrated in the case of the jumper, it is perfectly possible for a force to *accelerate* a system without putting any work into it.

The work-KE theorem, derived from Newton's second law, is an extremely restricted relation. It hints at the energy concepts and guides us into articulation of the first law of thermodynamics as an independent statement about order in nature, but it is not, in itself, a general statement about energy conservation. We must continue with our *discovery* of how to calculate the numbers that obey the general conservation law.

Question 4.21.3 Set up the FLT equation for what happens in the earth-jumper system after the instant the feet of the jumper lose contact with the ground. Note that you will now have to change various symbols, adopting ones appropriate to the succeeding story. Is there any appreciable heat transfer Q ? Is any external work done on the system? Which internal energy changes are significant? Which are negligible?

Question 4.21.4 Suppose you stand on roller skates on a horizontal floor and push yourself away from a wall. For the time being let us still consider friction to be negligible. Write down the applicable cm equation (corresponding to Eq. 4.21.2 above). Defining the system under consideration carefully, write down the relevant FLT story corresponding to Eqs. 4.21.6 and 4.21.7. Explain steps in your own words and interpret the final result. Does the wall do work on your body? What is the source of the KE acquired by your body? Is there any change in PE? Note that, like the case of the jumper, we have here an instance in which a force accelerates a deformable system but puts zero work into it. The energy transformations leading to the kinetic energy of the system are entirely internal.

Question 4.21.5 Discuss various phases of jumping on a trampoline from the points of view we have developed above. (This is an open ended question, and there is no one simple pat answer. Make it your own story; it can help deepen your insight into the new concepts.)

4.22 WORK AND HEAT IN THE PRESENCE OF SLIDING FRICTION

Consider the familiar physical situation represented in Fig. 4.22.1: A block of mass m is accelerated along a horizontal floor or table by an applied force of magnitude P against a resisting force of sliding friction of magnitude f .

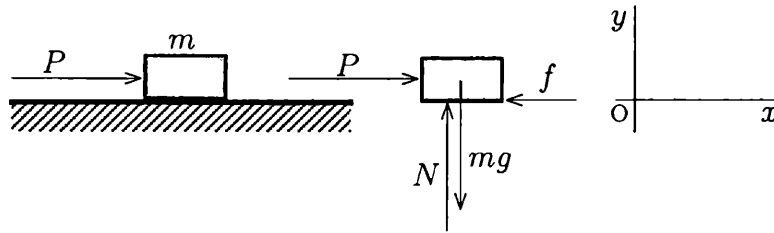


Figure 4.22.1 A block of mass m is accelerated along a horizontal surface by a force of magnitude P against the opposing force of sliding friction of magnitude f .

We start in by showing that one runs into an anomaly or paradox in describing energy transformations in this situation unless we are very careful about our use of the concept of “work.” Suppose that we consider the simplest case in which we keep the block moving at uniform velocity. The horizontal forces must then be equal in magnitude and opposite in direction, and the kinetic energy change is zero. The block is not in contact with any higher or lower temperature system, so no heat is being transferred, i.e., $Q = 0$. We then have, from the work-KE equation Eq. 4.19.3 applied to the block as the system under consideration:

$$(P - f)\Delta x_{\text{cm}} = 0 \quad (4.22.1)$$

or

$$P\Delta x_{\text{cm}} = f\Delta x_{\text{cm}} \quad (4.22.2)$$

If we take this statement, coming from the center of mass equation rather than from the FLT, as a statement about quantities of work, it says that the net work done on the system (taken to be the block) is zero. [That is, the work $P\Delta x_{\text{cm}}$ put *into* the system (block) by the force P and the work $f\Delta x_{\text{cm}}$ taken *out* of the system under action of the force f , are equal in magnitude and add up to zero.] If there is zero net work done on the system and zero input of heat Q from the outside, there cannot be any internal energy change of the system and hence no temperature rise. Yet we know that, in the presence of friction, there is a temperature rise not only in the block but also in the floor along which the block is rubbing. There is clearly something wrong in describing both sides of Eq. 4.22.2 as quantities of work done on and by the

system if we want to find those formulas and connections that keep giving us conservation of quantities of energy. The following analysis will ferret out where we go wrong in the preceding story.

In the preceding analysis, the quantity $P\Delta x_{\text{cm}}$ is a real amount of work done on the block since it affects the entire block as though the latter were a single particle: Every chunk of the block undergoes exactly the same displacement. The difficulty cited in the preceding paragraph arises in the fact that, although $P\Delta x_{\text{cm}}$ is a real quantity of work done on the block, the quantity $f\Delta x_{\text{cm}}$ is *not* a real work done by the block on its surroundings. The latter, as we shall see, is what might be called “pseudo work.” The real work done at the block-floor interface does not involve the whole body displacement Δx_{cm} but involves a complicated mess of small displacements of material at the interacting surfaces of the block and the floor under the influence of the smeared out (rather than concentrated) force f . We are completely unable to describe these displacements quantitatively in any simple or direct way.⁹ Furthermore, real work is being done in this messy way not just on the block but also on the floor, and, with different kinds of material in each object, we do not know how the displacements are distributed. We shall return to this point later. First let us apply the FLT:

$$\Delta E \equiv \Delta E_{\text{therm}} + \Delta E_{\text{chem}} + \Delta E_{\text{kin}} + \Delta E_{\text{pot}} + \Delta E_{\text{misc}} + \dots = Q + W \quad (4.22.3)$$

and see how it eliminates the paradox.

Since the difficulty we are encountering resides in the fact that we are unable to describe the work done on each interacting body at the interface of the block and the floor, the thing to do is choose a system in which we do not have to deal with this problem, i.e., let us find a way of sweeping the problem under the rug. The simplest system to take is the combination of block and floor. The complicated interaction at the interface is then hidden within the system, and we do not have to analyze its details. (The experience we are now having dramatizes the fact that the choice of system in our energy analyses is entirely up to us and that it is perfectly legitimate to use this freedom of choice to simplify the problem.)

Let us now consider the more general case in which $P > f$, and the block is being accelerated from rest. The cm Eq. 4.19.3 applied to the block then tells us that

$$(P - f)\Delta x_{\text{cm}} = \frac{1}{2}mv_{\text{cm}}^2 - 0 \quad (4.22.4)$$

or

$$P\Delta x_{\text{cm}} = f\Delta x_{\text{cm}} + \frac{1}{2}mv_{\text{cm}}^2 \quad (4.22.5)$$

⁹Sherwood, B. A., and Bernard, W. H. “Work and Heat Transfer in the Presence of Sliding Friction,” *Am. J. Phys.* **52**, 1001 (1984) develop a plausible model for displacements at the interface under certain conditions. An interested reader will find the discussion illuminating.

where $v_{\text{cm f}}$ denotes the final velocity at the instant of cm displacement Δx_{cm} .

Now let us apply the FLT [Eq. 4.22.3] to the block-floor system, excluding the surrounding air: The only work being done on this system is $P\Delta x_{\text{cm}}$. Therefore $W = P\Delta x_{\text{cm}}$. There is zero heat transfer Q except for a possible small loss of heat to the surrounding air as the temperature of the system increases with increasing thermal internal energy. We shall retain the symbol Q , remembering that it stands only for this small *loss* (a negative quantity) and not for any transfer of heat *into* the system. The internal energy changes are only thermal and kinetic translational, there being no rotational effects, no chemical effects, and no storage of PE. We therefore obtain

$$\Delta E_{\text{therm}} + \frac{1}{2}mv_{\text{cm f}}^2 = P\Delta x_{\text{cm}} + Q \quad (4.22.6)$$

or, if we take the heat loss Q to be negligible,

$$\Delta E_{\text{therm}} = P\Delta x_{\text{cm}} - \frac{1}{2}mv_{\text{cm f}}^2 \quad (4.22.7)$$

Interpretation of Eq. 4.22.7 yields the following statements:

- (1) The increase in the thermal internal energy of the block-floor system is directly equal to that part of the total work $P\Delta x_{\text{cm}}$, done by the force P , that does not go into the form of translational KE of the block (providing that a negligible amount of heat is transferred to the surrounding air).
- (2) This amount of work, which is said to be “dissipated,” is directly transformed into thermal internal energy of both the block and the floor (we are unable to apportion the distribution between the two bodies), leading to the observable temperature increase in both bodies.
- (3) The temperature rise and increase in thermal internal energy of the system do *not* result from a transfer of heat Q to the system from the surroundings.
- (4) The quantity $f\Delta x_{\text{cm}}$ does not appear anywhere in our energy equation; it is not a quantity of work put into or taken out of the system.
- (5) If P and f happen to be equal in magnitude and the block moves at uniform velocity without change in KE, $P\Delta x_{\text{cm}}$ and $f\Delta x_{\text{cm}}$ happen to be equal in magnitude, but that does not make the two quantities identical conceptually, and it does not make the net work done on the system equal to zero.

These statements resolve the paradox enunciated earlier in our first encounter with Eq. 4.22.2.

You may have heard it said (and you may have used the expression yourself) that, in the presence of friction, such as in the situation now being discussed, “work is converted into heat.” In the light of the analyses we have

conducted and the language we have now developed, this is an unfortunate way of describing what happens in the observed energy transformations because it implies that heat is put into the system to increase the temperature when no such transfer has actually taken place at all. If anything, some heat is lost to the surrounding air as the temperature of the system increases. It is much better and clearer to say that, “when work is dissipated in frictional processes, it is directly converted into thermal internal energy of the system in which the dissipation occurs.” The temperature increase associated with this increase in thermal internal energy is that which *would have* taken place if an amount of heat, equivalent to the work dissipated, had been transferred to the system. It is this *would have . . . if . . .* aspect that tempts one to speak of “work being converted into heat,” but it is better to avoid this usage.

Now, having applied the FLT to the block-floor combination and avoided considering what happens to each object separately, let us see what we might be able to say about each object. The difficulty we have been avoiding resides in what happens at the interface—the rubbing of the two surfaces against each other.

The complexity of the atomic-molecular interactions that take place at surfaces in sliding contact with each other has been a subject of intensive research for many years because friction is universally present, and its reduction or minimization is very important in most engineering applications where motion and contact are involved between wearing surfaces. An indication, based on very modern techniques, of what happens to atoms and molecules at points of contact between different materials is displayed in a paper in *Science*.¹⁰ At small regions of close “contact,” between the surfaces, atoms or molecules of one surface migrate into the other surface and interpenetrate among the other molecules, causing a kind of “welding” in the small region. As one object slides over the other, these regions of high attraction are repeatedly “sheared off,” broken, and re-established. This, in somewhat over-simplified presentation, is a description of what leads to the frictional force we have denoted by f .

Fig. 4.22.2 illustrates what happens to layers of material near the interface between the block and the table. In the absence of the frictional force, the material near the interface is undeformed as in (a). When the material near the interface is subjected to the frictional forces parallel to the interface, the material is deformed in such a way that its layers shift sideways as in (b). This is called a “shear” deformation. Note that such a deformation can only be sustained in solid materials. In a liquid the layers immediately slide over each other so that there is no deformation or change in shape. (This does not mean that there is zero friction as layers of liquid slide over each other; there is still a rubbing effect of layer on layer which is given the name “viscous friction” or “viscosity.”) Only solids, however, can sustain a shear deformation.

¹⁰Landman, U., Luedtke, W. D., Burnham, N. A., and Colton, R. J., “Atomistic Mechanisms and Dynamics of Adhesion, Nanoindentation, and Fracture, *Science* **248**, 454 (1990).

You can easily make a model of what is shown happening in Fig. 4.22.2 by subjecting your book to forces as shown in (b) and seeing the pages of the book being displaced over each other. Another model is afforded by deforming a coil spring transversely as in Fig. 4.22.2.

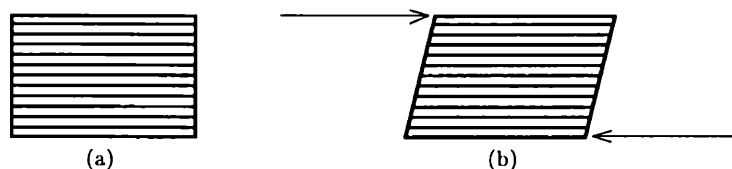


Figure 4.22.2 (a) Layers of material within a solid object *not* subject to external forces. (b) Layers of material within a solid object subject to forces in the direction indicated are displaced laterally. The deformation shown is called “shear.”

Now consider the action of the frictional force f in Fig. 4.22.1. This force is not concentrated in a small region as is our push on the block; it is “smeared out” over the entire interface between the block and the floor. The force f is the sum (or integral) of the smaller forces distributed over the entire area of the interface. Thus there is no single, concentrated force f that undergoes a displacement Δx_{cm} . What happens to the floor at the interface, for example, is that the small local forces (that add up to f) produce small local shear displacements of the layers of the floor at regions where the surfaces are especially close and are “welded” together. As the block slides, these regions break apart and reform, sending vibrations through the solid materials. The net effect is that work is being done on the floor as the sum of the small local forces multiplied by their associated small displacements, but the sum of these small work quantities is *not* $f\Delta x_{\text{cm}}$; in fact we cannot calculate this amount of work directly because we have no numerical information as to the magnitudes involved. (We shall, however, say something about the sums from another point of view shortly.) In any case, these work quantities, put into the floor at the interface, are converted into internal thermal energy of the floor by a succession of effects that eventually reach down to the atomic-molecular level. The frictional forces acting in the opposite direction on the block have the same overall effect on the block, the quantities of work distributed over the bottom of the block being converted into internal thermal energy of the block.

Let us describe in symbols the overall effects we have just outlined. Let W_F represent the work done by the frictional effects on the floor, and let W_B represent the work done by the frictional effects on the block. As we pointed out above, we cannot calculate either W_F or W_B from fundamental principles, but we know that, if the block is sliding at uniform velocity (without increase in KE), the two quantities must add up to $P\Delta x_{\text{cm}}$, since that is the total amount of work dissipated, i.e.,

$$W_F + W_B = P\Delta x_{\text{cm}} \quad (4.22.8)$$

Therefore

$$W_B = P\Delta x_{cm} - W_F \quad (4.22.9)$$

Let us now apply the FLT to the block and the floor as separate systems. Taking heat transfer Q to be negligible for each system, using the values of W indicated above, and taking the case of uniform velocity of the block (i.e., no change in KE), we obtain the following two equations. For the floor:

$$\Delta E_{\text{therm } F} = W_F \quad (4.22.10)$$

and for the block:

$$\Delta E_{\text{therm } B} = P\Delta x_{cm} - W_F \quad (4.22.11)$$

These are the thermal internal energy changes for each body separately. W_F remains an unknown quantity, but we can see the nature of the distribution between the floor and the block. If these objects are composed of identical material, it is reasonable to suppose that W_F and W_B are equal and that the work dissipated is equally distributed between the two objects. Under other circumstances the distribution need not be equal. If we add Eqs. 4.22.10 and 4.22.11 we return to the situation of the combined system of block and floor and recover Eq. 4.22.7 (without the KE term.)

Question 4.22.1 In a manner parallel to that of the preceding discussion, analyze the case in which the block is initially sliding along a floor with initial velocity $v_{cm i}$ and coasts to a stop under the influence of the frictional force f . The force P is zero throughout this sequence of events. (In this instance it is, of course, KE that is completely dissipated rather than work.) Draw the relevant force diagram. Write down the cm equation. Then apply the FLT and interpret what both equations tell us. Be sure to use the formalism rigorously, being especially careful about all the algebraic signs. [Note, for example, that the change in KE (final minus initial value) is a negative and not a positive quantity.] Now apply the work-KE theorem (i.e., the cm equation) and obtain an expression for how far the block coasts against the frictional force f before coming to a stop.

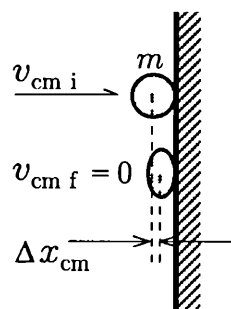
Question 4.22.2 Analyze the case of a car accelerating along a horizontal road: Draw the force diagram for the car. (Note that the total driving force f_D imparting acceleration to the car is the sum of the frictional forces exerted by the road on the tires of the driving wheels.¹¹ The driving wheels may be the two rear wheels, or the two front wheels, or all four wheels, depending on the

¹¹It is sometimes said, rather carelessly, that “friction always opposes motion.” We must phrase our statements about such matters much more carefully. It is true that a frictional force always opposes slipping of one surface over another. That is the correct sense in which it “opposes motion.” In the case of acceleration of the car, however, the frictional force exerted by the road on the tires of the driving wheels is imparting acceleration to the car and not opposing the motion of this object (the car) even though it *does* oppose the slipping of the tires at the road surface.

car design.) Let f_R denote the sum of all the frictional forces resisting the motion, such as air resistance and friction in various bearings. Suppose the car starts from rest and attains a velocity $v_{cm f}$ after a displacement Δx_{cm} . Write the cm equation. Apply the FLT, keeping a term for the various internal kinetic energies other than the translational KE of the whole car. Interpret the results. (Note that f_D is *not* a work doing force even though it does impart acceleration to the car. The situation here is quite analogous to the effect of the normal force in the cases in which you jumped vertically upwards and pushed yourself away from the wall while standing on roller skates.) Where do the KE of the car and the energy dissipated through frictional effects come from?

Question 4.22.3 Suppose we throw a ball of mass m horizontally against a vertical wall as shown in Fig. 4.22.3. (We shall ignore vertical effects.) We assume the ball is not spinning and that it has velocity $v_{cm i}$ at the instant it makes contact with the wall. The cm of the ball continues to be displaced toward the wall as the ball flattens and deforms until the cm velocity becomes zero (i.e., just before the rebound would begin.) The wall exerts an average normal force \bar{N} on the ball during the interval just described.

Figure 4.22.3 Ball of mass m , thrown horizontally, collides with a rigid wall with velocity $v_{cm i}$. As the ball deforms on collision, the cm displacement is Δx_{cm} at the instant $v_{cm} = 0$, and the rebound is about to begin.



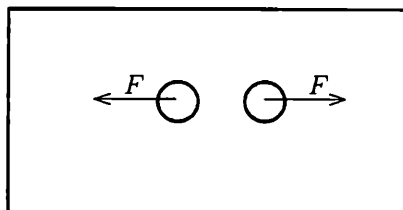
- Draw a force diagram for the ball, and write the cm equation for the ball for the interval described, being very careful about algebraic signs.
- (Note that the normal force \bar{N} is a zero work force: Although the ball deforms, the displacement of the material of the ball right at the wall is parallel to the wall and therefore perpendicular to the normal force.) Apply the FLT to the ball-wall system for the interval described, noting that some energy might be dissipated and some stored as PE in deformations. Interpret your results. What form does the FLT relation take and what will happen after the interval described if zero PE is stored (i.e., if the ball is made of putty and the collision is perfectly inelastic)? What form does the FLT relation take and what will happen after the interval described if there is zero dissipation (i.e., if the collision is perfectly elastic)?

Question 4.22.4 Consider a system consisting of two identical frictionless pucks of mass m on an air table as shown in Fig. 4.22.4.

Suppose we exert forces of equal magnitude F on each puck, acting in exactly opposite directions, with each puck starting from rest. Each puck is displaced

a distance Δx and acquires a final velocity v_f . (In the following analysis you must distinguish very carefully between the displacements Δx of the individual pucks and the displacement Δx_{cm} of the system of the two pucks. You must also distinguish very carefully between the individual velocities v_f of the individual pucks and the velocity of the center of mass v_{cm} of the two-puck system.)

Figure 4.22.4 Looking down on an air table. Two identical frictionless pucks are accelerated in opposite directions by forces of equal magnitude F .



- (a) Write the cm equation for this system, treating the velocity of the center of mass as an unknown and solve for this velocity. (You should get the result $v_{cm} = 0$. Does this make sense? Why or why not?)
- (b) Apply the FLT to the two puck system, noting that the forces F are both doing work on the system. Compare what happens in this case with what happened when you compressed a spring by displacing both ends instead of just one end. What are the similarities and what are the differences?

Question 4.22.5 By applying the FLT, analyze Joule's experiment, described in Sect. 4.14, in which the "mechanical equivalent of heat" was measured by having descending weights turn the paddle wheel in a vessel of water. Be sure to state explicitly what you define to be your system.

4.23 LOGICAL STATUS OF THE CONSERVATION LAWS

With slowly growing acceptance of the principle of conservation of energy enunciated by Mayer and Joule, the concept was extended quantitatively to include other mechanical phenomena such as wave propagation and the flow of fluids. The young German physicist Helmholtz (1821-1894) published a paper in 1847 in which he gave a quantitative treatment of the energetics of certain simple electric, magnetic, and chemical phenomena. Conservation of energy played an important role in the development of an understanding of light and electromagnetism; it is one of the fundamental pillars of the modern theory of relativity.

When, during the 1930s, phenomena observed in certain types of radioactive decay (beta emission) seemed to violate the laws of conservation of energy and momentum, faith in the conservation laws had become so great that it was postulated that a virtually unobservable particle with zero electrical charge and zero rest mass, called the "neutrino," was emitted and that this elusive entity accounted for the missing linear momentum, angular momentum, and energy. Later this same hypothetical entity proved useful in explaining other

nuclear phenomena, and considerable faith was built up in the construct because of its wide usefulness and serviceability in preserving the conservation laws even though no interactions between neutrinos and other known particles had been observed. It was predicted that neutrinos should occasionally interact with atoms among which they passed, but such interactions were estimated to be so exceedingly infrequent that detection seemed virtually impossible. It was not until 1952 that experimental techniques, sufficiently sensitive and sophisticated to detect such rare events, were developed. The search proved successful, and the predicted interactions with atoms were observed, verifying the existence of neutrinos. Frederick Reines shared the 1955 Nobel Prize in physics for his work in 1952. At present, study of neutrinos, their interactions with atoms, and their cosmic origins are major areas of research at the cutting edge of contemporary physics.

As a result of such experiences, built up over a period of over 150 years, scientists have developed a profound faith in the fundamental validity of the conservation laws as expressions of order in nature. It is therefore desirable to scrutinize a little more deeply the logical status of these assertions. Are the conservation laws essentially convenient definitions or conventions, as the great French mathematician and physicist Poincaré implied in some of his writings early in the 20th century?¹² If the conservation law for energy appears to fail in some new investigation, will we always be able to rehabilitate it by inventing a new particle or a new form of energy?

Professor Eric Rogers beautifully illustrates some of the points at issue in the following dialog in which "You" and "Faustus" debate a theory of the origin of frictional forces:¹³

You. I don't believe in demons.

Faustus. I do.

You. Anyway, I don't see how demons can make friction.

Faustus. They just stand in front of things and push to stop them from moving.

You. I can't see any demons even on the roughest table.

Faustus. They are too small, also transparent.

Y. But there is more friction on rough surfaces.

F. More demons.

Y. Oil helps.

F. Oil drowns demons.

Y. If I polish the table, there is less friction and the ball rolls farther.

F. You are wiping the demons off; there are fewer to push.

Y. A heavier ball experiences more friction.

¹²See Chapters 6 and 7 of "Science and Hypothesis" by Henri Poincaré. Dover Publications, New York, 1952.

¹³From "Physics for the Inquiring Mind" by Eric Rogers, Princeton University Press, 1960.

- F. More demons push it, and it crushes their bones more.
- Y. If I put a rough brick on a table, I can push against friction with more and more force, up to a limit, and the block stays still, with friction just balancing my push.
- F. Of course, the demons push just hard enough to stop you from moving the brick; but there is a limit to their strength, beyond which they collapse.
- Y. But when I push hard enough and get the brick moving, there is friction that drags the brick as it moves along.
- F. Yes, once they have collapsed, the demons are crushed by the brick. It is their crackling bones that oppose the sliding.
- Y. I cannot feel them.
- F. Rub your finger along the table.
- Y. Friction follows definite laws. For example, experiment shows that a brick sliding along the table is dragged by friction with a force independent of velocity.
- F. Of course, same number of demons to crush however fast you run over them.
- Y. If I slide a brick along the table again and again, the friction is the same each time. Demons would be crushed on the first trip.
- F. Yes, but they multiply incredibly fast.
- Y. There are other laws of friction. For example, the drag is proportional to the pressure holding the surfaces together.
- F. The demons live in the pores of the surfaces. More pressure makes more of them rush out to push and be crushed. . . .

If matters are kept on this plane, and no connection is made to other physical phenomena, knowledge, or properties, the term “friction” is nothing more than a name synonymous with the behavior of Faustus’s conspiratorial society of demons. Faustus’s explanations are completely *ad hoc*—in other words, they are expressly concocted to cover each particular point; there is no way of refuting them. If we invoke some new observation or experiment in an attempt to test a particular statement and show it to be false, Faustus will invent an appropriate demonic activity to cover the new case. Such a model or hypothesis is said to be “unfalsifiable,” meaning that it is impossible to refute it by appeal to experience or experiment.

If we repeatedly extend the conservation laws to incorporate newly discovered phenomena, it is legitimate to ask whether “momentum” and “energy” are names for demons. Are the conservation laws *ad hoc*, unfalsifiable statements, adopted by convention and used because of their simplicity and convenience?

Karl Popper, an eminent logician and philosopher of science, argues¹⁴ that scientific hypotheses can never be conclusively verified or “proved” because

¹⁴ “The Logic of Scientific Discovery,” Basic Books, Inc., New York., 1959.

it is impossible to test them on each of the infinity of particular cases to which they might apply, but he points out that scientific hypotheses might at least be distinguished from mathematical or even metaphysical systems, by the criterion of falsifiability. It should be possible, in principle, to refute and prove them false by appeal to experience; just one failure would do the trick:

. . . what characterizes the empirical method is its manner of exposing to falsification, in every conceivable way, the system to be tested. Its aim is not to save the lives of untenable systems, but, on the contrary, to select the one which is by comparison the fittest, by exposing them all to the fiercest struggle for survival.

And P. W. Bridgman, philosopher of science and Nobel laureate for research in high pressure phenomena, argued that the conservation laws are indeed falsifiable in this sense and that they are far from tautologies or pure conventions:¹⁵

A remark of Poincaré is often quoted to the effect that if we ever found the conservation law for energy appearing to fail we would recover it by inventing a new form of energy. This seems to me to be a misleadingly partial characterization of the situation. If in any specific situation the law apparently failed, we would doubtless first try to maintain the law by inventing a new form of energy, but when we had invented it we would demand that it be a function [of numbers, or parameters, that describe the state of the system] and that the law would continue to hold for all the infinite variety of combinations into which the new parameters might be made to enter. Whether conservation would continue to hold under such extended conditions, could be determined only by experiment. The energy concept is far from being merely a convention.

Bridgman might have added that the law of conservation of energy is also falsifiable through the concept of perpetual motion machines of the kind described in earlier sections. Construction of a machine that does indeed deliver work to external systems indefinitely without change of final state of the machine would make the conservation law untenable.

Our experience to date with the laws of conservation of momentum, energy, and mass does indeed satisfy Bridgman's criterion. It is the far flung network of successfully achieved experimental and theoretical connections, interlinkages, cross-checks—the entire fabric of which can be tested for internal consistency—that encourages us to believe that the Newtonian synthesis and the conservation principles are something more than arbitrary definitions or demonologies.

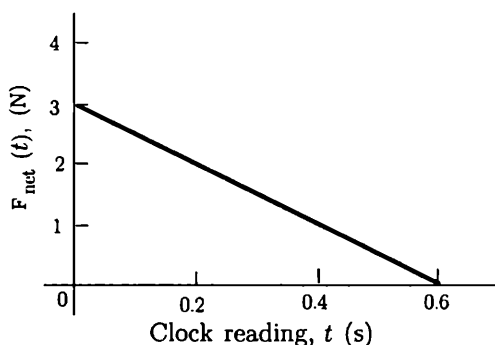
¹⁵ "The Nature of Thermodynamics," Harvard University Press, 1941.

4.24 QUESTIONS AND PROBLEMS

4.24.1 Being sure to define the system you choose to consider, describe, both in words and by substitution in the FLT expression:

- the energy transformations that are taking place while you keep walking at constant speed.
- the energy transformations that are taking place during an interval in which you are running at increasing speed up a slope.

4.24.2 The net force acting in the positive direction horizontally on a glider on a level air track (essentially frictionless system) varies with time as shown in the diagram. The glider has a mass of 0.850 kg. When the force is applied to the glider abruptly at $t = 0.00$ s, the glider has an instantaneous velocity of +0.150 m/s.



- Describe in your own words what happens to the glider over the time interval between $t = 0.00$ and $t = 0.60$ s. Use the vocabulary of impulse, momentum, work, and kinetic energy rather than that of Newton's second law.

Perform the following calculations using the concepts and relations referred to in your verbal description. In each calculation describe your reasoning briefly.

- Calculate the momentum of the glider at $t = 0.00$ s (i.e., the initial momentum of the glider.)
- Calculate the net impulse delivered to the glider by the applied net force between $t = 0.00$ and $t = 0.60$ s.
- Calculate the *change* in momentum of the glider over this time interval.
- Calculate the momentum of the glider at clock reading $t = 0.60$ s (i.e., the final momentum of the glider.)
- Calculate the instantaneous velocity of the glider at clock reading $t = 0.60$ s.
- Calculate the initial and final values of the kinetic energy of the glider.
- Calculate the *change* in kinetic energy of the glider.
- Calculate the work that must have been done by the net force in order to produce this change in kinetic energy.
- Calculate the potential energy change of the glider-earth system over the interval under consideration.

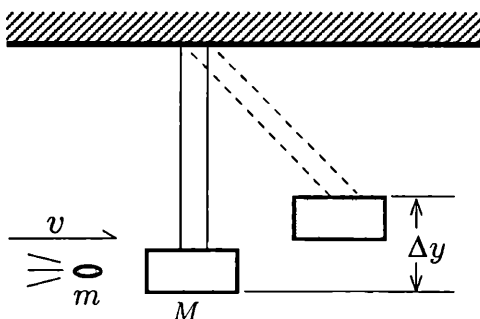
4.24.3 Note to the instructor: For a problem like 2.24.2 but with position rather than clock reading as the independent variable, see Problem 2 at the end of Chapter 5 in Part I.

4.24.4 Return to the example discussed in Sect. 4.11, that of throwing a ball vertically upward. Assume a reasonable numerical value for the mass of the ball, a reasonable numerical value for the average force you might exert with your hand, and a reasonable value for the vertical displacement of your hand during the act of throwing. Then proceed to obtain numerical values for each of the various quantities and changes described verbally in the example. Revise your choice of numerical values if your results turn out to be unreasonable, for example, if the vertical rise turns out to be absurdly high or absurdly low.

4.24.5 A simple pendulum consists of a bob of mass $m = 0.543$ kg and a string of length $L = 85$ cm. The bob is released from rest at the same level as the suspension point, with the string taught. Using energy arguments and explaining your reasoning step by step, estimate the instantaneous velocity of the bob as it passes through the bottom of its swing. Obtain an algebraic expression for this velocity first, and then put in relevant numbers. How does your result compare with the velocity the bob would attain on vertical free fall over the same interval of height?

4.24.6 A common method of determining the velocity of a bullet is to use the system known as a “ballistic pendulum,” shown in the figure.

The bullet is fired into a massive block suspended by wires and remains embedded in the block. The height Δy through which the block-bullet combination rises after impact is measured. Knowing the masses m and M of the bullet and block, respectively, one can calculate the original velocity v of the bullet.



- (a) Confirm this statement by analyzing the problem algebraically, noting that there are two successive stages involved. The first stage is the perfectly inelastic collision in which the bullet-block combination takes on the instantaneous velocity v_c and to which momentum considerations apply. The second stage is the swing of the pendulum to which energy considerations apply. Analyze this latter stage in the symbols and language of the FLT, showing that

$$v_c = \left(\frac{m}{m + M} \right) v$$

and that

$$v = \left(\frac{m + M}{m} \right) \sqrt{2g\Delta y}$$

- (b) Suppose you wish to set up an experiment in which you will use bullets known to have a mass of about 10 g and a velocity of the order of 800 m/s. Design the rest of the experiment: What mass M would you elect to use for the pendulum block and what rise Δy would you expect? Be sure you are dealing with reasonable values.
- (c) Examine what happens to the kinetic energy of the bullet. Is it conserved or dissipated? Show that the ratio of the initial KE of the bullet to the initial KE of the bullet-block combination immediately after collision reduces to the simple expression $[m/(m + M)]$. Interpret this result. What happens to KE in such collisions if $m = M$, if $m \ll M$, if $m \gg M$?

4.24.7 Using the symbols and language of the FLT, calculate how much frictional work must be supplied in order to boil away 1500 g of water initially at 20°C. What assumptions are implied in your calculation? Explain your reasoning. Compare your result with other work quantities you can estimate. With respect to what situations is this a large amount of work? With respect to what situations is it small?

4.24.8 Using the symbols and language of the FLT and making use of relevant information from Table 3.5.1, estimate the lower limit of velocity that a lead bullet must have in order to become melted when stopped by a perfectly inelastic collision with a rigid wall. Assume the initial temperature to be some ordinary atmospheric value. What idealizations are implied in your calculation? Why is your answer a lower rather than an upper limit? Why is no specification made of the mass of the bullet? Actually, the initial temperature of the bullet just before striking the wall is likely to be considerably higher than the ambient (surrounding) atmospheric temperature. Why? What influence does this have on your estimate? (Ans. Of the order of 350 m/s.)

Index

A

Action at a distance, 36

C

Caloric theory, 115

Rumford's attack on, 116

Calorie, 72

Calorimetry, 72ff

Center of mass, 37ff

Classical physics, 20, 36

Collisions, 2ff, 87ff

classification of, 5

conservation of momentum in, 15

elastic, 5, 84, 88

inelastic, 5, 84, 88, 110ff

partly elastic, 6

Conservation, 62

of mass, 63ff

Conservation laws, 83ff

and non-spontaneous changes, 85

energy, 86, 119

logical status of, 144ff

mechanical energy, 105

D

Deductive reasoning, 14, 83

Density, 50, 65

Dynamics, 1, 47

E

Energy, 83ff

as an abstract construct, 104

conservation of, 86, 102ff, 119, 121

and extended systems, 123ff, 129ff

and first law of thermodynamics, 124ff

mechanical energy, 105

dimensions of, 109

dissipation of, 105

forms of, 86

internal, 126ff

varieties of, 127ff

kinetic, 89, 94, 103ff

meaning of "energy", 86

potential, 104ff

thermal, 125

transformations of, 86, 92, 102ff

in horizontal displacement, 89ff

in vertical displacement, 94ff, 133ff

with Hooke's law spring, 99ff, 130ff

units of, 109

using the vocabulary of, 111ff

work, 102ff

Extended bodies, 129ff

and energy conservation, 123ff

versus point masses, 123ff

work and deformation of, 124ff

F

Falsifiability of theories, 146

First law of thermodynamics, 124ff

Force

conservative force, 105, 108

path independence, 106, 108

dissipative force, 105, 108

path dependence, 108

time average, 24

Frames of reference, 8ff

Friction

and dissipation of work, 124

work and heat in presence of, 137ff

H

Heat, 67ff, 105, 124

algebraic signs of, 125

and work in presence of friction, 137ff

as an imponderable fluid, 114

caloric theory of, 115

Joule's experiments with, 119ff

latent, 72ff

measuring quantity of, 70
 models of, 114ff
 not a function of state, 80, 125
 path dependence of, 80
 relation to work, 118
 released in the boring of cannon,
 117, 120
 transfer of, 69
 weight of, 117

I

Imponderable fluids, 114
 Impulse, 22ff, 87ff
 and average force, 24
 and momentum, 22ff
 net impulse, 26
 Impulse-momentum theorem, 26
 Inductive reasoning, 14, 83, 124
 Inertia, law of, 37
 Integral
 path dependent, 23, 108
 path independent 108
 Interaction, 45
 chemical, 46
 electrical, 46
 electromagnetic, 46
 gravitational, 46
 magnetic, 46
 mechanical, 45
 nuclear, 46
 thermal, 45
 Internal energy, 126, 127ff
 Invariance, 19ff

K

Kinematics, 1, 47
 Kinetic energy, 103ff

L

Latent heat, 77ff
 Lavoisier and conservation of mass, 63ff
 Laws of nature, 84

M

Mass, conservation of, 63, 83
 Method of mixtures, 72
 Momentum, 2, 15ff, 82ff
 and net impulse, 26
 conservation of, 16, 19, 24, 30, 33ff,
 36, 83

Motion, quantity of, 15

N

Newton's second law
 integral with respect to position
 constant force, 89ff
 varying force, 96ff
 Hooke's law force, 99ff
 integral with respect to time, 20ff,
 87ff
 Newton's third law and momentum con-
 servation, 31ff
 Neutrino, 144

P

Pascal's law, 56
 Perpetual motion, 85
 and falsifiability of energy concept,
 147
 Phase, change of, 69, 77ff
 Phlogiston, 114
 Potential energy, 104ff
 algebraic signs of, 105
 as a property of a system, 106
 forms of, 105
 path independence of, 106
 zero reference level for, 108
 Pressure, 50, 51ff
 atmospheric, 57
 in fluids, 54ff
 Properties of a system, 46, 48

R

Relativistic phenomena, 19, 20
 Relativity, 8, 18ff, 36
 principle of, 18ff
 Restricting scope of inquiry, 1ff

S

Shear, 56
 Specific heat, 72, 75
 as a varying quantity, 76
 Spontaneous change, 85
 State, 48ff
 change in, 49, 124ff
 functions of, 49, 80
 of a system, 48
 variables of, 49ff
 extensive, 50
 intensive, 50

System, 17
 closed, 17, 26
 open, 17, 26
 properties of, 48
 state of, 48
 subsystem, 47

T

Temperature, 59
 changes without heat transfer, 81
 distinguished from "heat," 67
 scales of, 60
 Thermal equilibrium, 60
 Thermometer, 59

V

Variables (of state)
 extensive, 50
 independent, 50
 intensive, 50
 Viscosity, 57
 Vis viva, 87

W

Work, 89, 94, 102ff, 124
 algebraic signs of, 103, 125, 126
 and frictional dissipation, 123, 137ff
 and deformation of extended bodies, 123
 not a function of state, 125
 relation to heat, 118
 Work-energy theorem, 104

Z

Zero-work forces, 103

